

Methods in
Molecular Biology 1558

Springer Protocols

Cathy H. Wu · Cecilia N. Arighi
Karen E. Ross *Editors*

Protein Bioinformatics

From Protein Modifications
and Networks to Proteomics

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
<http://www.springer.com/series/7651>

Protein Bioinformatics

From Protein Modifications and Networks to Proteomics

Edited by

Cathy H. Wu

Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA

Cecilia N. Arighi

Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA

Karen E. Ross

*Department of Biochemistry and Molecular and Cellular Biology,
Georgetown University Medical Center, Washington, DC, USA*

 **Humana Press**

Editors

Cathy H. Wu
Center for Bioinformatics and Computational
Biology
University of Delaware
Newark, DE, USA

Cecilia N. Arighi
Center for Bioinformatics and Computational
Biology
University of Delaware
Newark, DE, USA

Karen E. Ross
Department of Biochemistry and Molecular and
Cellular Biology
Georgetown University Medical Center
Washington, DC, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-6781-0 ISBN 978-1-4939-6783-4 (eBook)
DOI 10.1007/978-1-4939-6783-4

Library of Congress Control Number: 2016963231

© Springer Science+Business Media LLC 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Proteins are the building blocks of life. Researchers now routinely generate genome-scale data with next-generation sequencing and proteomics technologies that require functional interpretation in the systems biology context. This volume aims to introduce the bioinformatics research methods for proteins—from protein sequence and structure, protein post-translational modifications (PTMs), protein-protein interactions (PPIs) and networks, to proteomics—and to serve as a definitive source of reference providing both the breadth and depth needed on the subject of *protein bioinformatics*. It is an updated edition to our previous Volume 694 “*Bioinformatics for Comparative Proteomics*” published in 2011, now with special emphasis on protein PTMs and networks.

The volume is organized into four parts: (1) basic framework and major resources for analysis of protein sequence, structure, and function; (2) bioinformatics approaches and resources for studies and functional analysis of protein PTMs, including curated PTM databases, text mining methods, and integrative predictive systems; (3) PPIs and protein networks, including tools for PPI prediction and approaches for the construction of PPI and PTM networks with applications in functional analysis and disease discovery; and (4) bioinformatics approaches in proteomics, including computational methods for mass spectrometry-based proteomics and integrative analysis for alternative splice isoforms, PTMs, and functional discovery.

Part I: Fundamentals of Protein Bioinformatics consists of five chapters covering the analysis of protein sequence, structure, and function.

Chapter 1 presents a comprehensive review (with categorization and description) of major protein bioinformatics databases and discusses the challenges and opportunities for developing next-generation databases to support data integration and data analytics in the Big Data era.

Chapter 2 introduces the functionality and data provided by the UniProt Knowledgebase and provides a practical guide for using UniProt resources from searching, entry retrieval, feature display, to ID mapping, sequence searching, and alignment with example use cases.

Chapter 3 describes the Protein Ontology (PRO) framework to represent protein classes, proteoforms, and protein complexes and provides a tutorial for searching, browsing, and visualizing PRO in hierarchical, network and sequence alignment views with several examples.

Chapter 4 describes the CATH-Gene3D online resource and its use to study protein domains and their evolutionary relationships, including comparing structures, recognizing homologs predicting domains, and classifying superfamilies into functional families.

Chapter 5 illustrates the basic process for structure-based virtual screening using publicly accessible tools and resources with examples of molecular docking of chemical compounds against the 3D structure of the target and selection of compounds for further biological evaluation.

Part II: Protein PTM (Post-translational Modification) Bioinformatics consists of five chapters covering approaches and resources for functional analysis of protein PTMs.

Chapter 6 presents major functionalities of the Plant Protein Phosphorylation DataBase (P3DB) and illustrates the bioinformatics analyses that can be performed on protein phosphorylation data, from querying a single phosphorylation site to browsing networks.

Chapter 7 outlines protocols for searching and navigating UniCarbKB, a comprehensive resource for mammalian glycoprotein and annotation data, to find information including glycan structures, attached glycoproteins, and specific metadata from publications.

Chapter 8 presents commonly used databases for single-nucleotide variation (SNVs) and PTMs and describes a broad approach that can be applied to many scenarios for studying the impact of nonsynonymous SNVs on PTM sites of human proteins.

Chapter 9 discusses analytical platforms for studying protein oxidation, describes computational tools to identify redox sensitive proteins, and examines roles of cysteine redox PTMs in drug pharmacology.

Chapter 10 demonstrates the use of the web-based text mining tools, RLIMS-P and eFIP, to extract and uncover information about protein phosphorylation (kinase-substrate-site information) and phosphorylation-dependent PPIs from the scientific literature.

Part III: Protein Network Bioinformatics consists of six chapters covering tools and approaches for predicting PPIs and for discovery using protein and PTM networks.

Chapter 11 describes the procedures to construct the Reactome FI (functional interaction) network by integrating multiple data sources, predicting FIs and constructing a FI database, and illustrates how to use ReactomeFIViz for network-based analysis of a gene list.

Chapter 12 reviews the principles and applications of PRISM, a PPI prediction tool, and discusses its extensions to discover the effect of single residue mutations, model large protein assemblies, and reconstruct large PPI networks or pathway maps.

Chapter 13 presents NDEx (Network Data Exchange)—an online commons for collaboration and publication of biological networks; it provides a technical overview of the NDEx platform and describes its applications and services, including CyNDEx, the NDEx App for Cytoscape.

Chapter 14 describes how to explore functional associations between protein modifications using PTMcode—a database of known and predicted functional associations between pairs of PTMs, including predicted crosstalk between PTMs in interacting proteins.

Chapter 15 presents a framework for studying PTMs in the context of protein interaction networks and complexes, including the use of the COMPLETE web-based tool for the analysis of PTM data at the level of protein complexes.

Chapter 16 illustrates the use of iPTMnet—a resource with PTM information from text mining, curated databases, and ontologies along with visualization tools for exploring PTM networks, PTM crosstalk, and PTM conservation across species—in several PTM scientific use cases.

Part IV: Proteomic Bioinformatics consists of six chapters covering computational methods for proteomics and integrative analysis of splice isoforms and PTMs for functional discovery.

Chapter 17 introduces the concepts and methods for protein identification from a tandem mass spectrometry dataset using protein sequence database search engines and discusses preparation of data files, selection of search parameters, and basic interpretation of the results.

Chapter 18 describes the analysis of intact protein spectra through deconvolution, deisotoping, and searching with ProSight Lite, a tool for the analysis of top-down mass spectrometry data, and discusses the iterative use of the tool to characterize proteoforms and discover PTM sites.

Chapter 19 describes the key steps and tools used to translate data-independent acquisition mass spectrometry (DIA-MS) proteomic data into knowledge, including library construction, protein quantification, network and functional enrichment analysis, and PTM analysis.

Chapter 20 presents the methods for annotation and characterization of splice isoforms, using I-TASSER for folding and functional predictions, and isoFunc, MisoMine, and Hisonet for isoform-level analyses of network and pathway-based functional predictions and PPIs.

Chapter 21 presents a data analysis workflow for quantitative investigation of PTMs from high-throughput mass spectrometry, illustrated with an example of an interpretation pipeline combining open-source software tools and R scripts for computational and statistical analysis.

Chapter 22 describes PhosphOrtholog—a web tool that maps annotated and novel orthologous PTM sites from high-throughput mass spectrometry data—and illustrates its use with examples of mapping novel PTM sites from cross-species phosphoproteomics data.

This volume is intended for readers who wish to learn about state-of-the-art bioinformatics databases and tools, novel computational methods, and future trends in protein and proteomic data analysis in systems biology. The audience may range from graduate students embarking upon a research project, to practicing biologists working on protein, proteomics, and systems biology research, and to bioinformaticians developing advanced databases, analysis tools, and integrative systems. With its interdisciplinary nature, the book is expected to find a broad audience in the biotechnology and pharmaceutical industries and in various academic departments in biological and medical sciences (such as biochemistry, molecular biology, protein chemistry, and genomics) and computational sciences and engineering (such as bioinformatics and computational biology, computer science, and biomedical engineering).

We thank our series editor Dr. John Walker for reviewing the chapter manuscripts and providing constructive comments. We thank all the authors and coauthors who contributed to the excellent scientific content for this volume.

Newark, DE, USA
Newark, DE, USA
Washington, DC, USA

Cathy H. Wu
Cecilia N. Arighi
Karen E. Ross

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>xi</i>

PART I FUNDAMENTALS OF PROTEIN BIOINFORMATICS

1 Protein Bioinformatics Databases and Resources	3
<i>Chuming Chen, Hongzhan Huang, and Cathy H. Wu</i>	
2 UniProt Protein Knowledgebase	41
<i>Sangya Pundir, Maria J. Martin, and Claire O'Donovan</i>	
3 Tutorial on Protein Ontology Resources	57
<i>Cecilia N. Arighi, Harold Drabkin, Karen R. Christie, Karen E. Ross, and Darren A. Natale</i>	
4 CATH-Gene3D: Generation of the Resource and Its Use in Obtaining Structural and Functional Annotations for Protein Sequences	79
<i>Natalie L. Dawson, Ian Sillitoe, Jonathan G. Lees, Su Datt Lam, and Christine A. Orengo</i>	
5 Structure-Based Virtual Screening	111
<i>Qingliang Li and Salim Shah</i>	

PART II PROTEIN PTM (POST-TRANSLATIONAL MODIFICATION) BIOINFORMATICS

6 Bioinformatics Analysis of Protein Phosphorylation in Plant Systems Biology Using P3DB	127
<i>Qiuming Yao and Dong Xu</i>	
7 Navigating the Glycome Space and Connecting the Glycoproteome	139
<i>Matthew P. Campbell, Robyn A. Peterson, Elisabeth Gasteiger, Julien Mariethoz, Frederique Lisacek, and Nicolle H. Packer</i>	
8 Impact of Nonsynonymous Single-Nucleotide Variations on Post-Translational Modification Sites in Human Proteins	159
<i>Naila Gulzar, Hayley Dingerdissen, Cheng Yan, and Raja Mazumder</i>	
9 Analysis of Cysteine Redox Post-Translational Modifications in Cell Biology and Drug Pharmacology	191
<i>Revati Wani and Brion W. Murray</i>	
10 Analysis of Protein Phosphorylation and Its Functional Impact on Protein-Protein Interactions via Text Mining of the Scientific Literature	213
<i>Qinghua Wang, Karen E. Ross, Hongzhan Huang, Jia Ren, Gang Li, K. Vijay-Shanker, Cathy H. Wu, and Cecilia N. Arighi</i>	

PART III PROTEIN NETWORK BIOINFORMATICS

- 11 Functional Interaction Network Construction and Analysis
for Disease Discovery 235
Guanming Wu and Robin Haw
- 12 Prediction of Protein Interactions by Structural Matching: Prediction
of PPI Networks and the Effects of Mutations on PPIs that Combines
Sequence and Structural Information 255
Nurcan Tuncbag, Ozlem Keskin, Ruth Nussinov, and Attila Gursoy
- 13 NDEx: A Community Resource for Sharing and Publishing
of Biological Networks 271
*Rudolf T. Pillich, Jing Chen, Vladimir Rynkov, David Welker,
and Dexter Pratt*
- 14 Bioinformatics Analysis of Functional Associations of PTMs 303
Pablo Minguéz and Peer Bork
- 15 Bioinformatics Analysis of PTM-Modified Protein Interaction
Networks and Complexes 321
Jonathan Woodsmith, Ulrich Stelzl, and Arunachalam Vinayagam
- 16 iPTMnet: Integrative Bioinformatics for Studying PTM Networks. 333
*Karen E. Ross, Hongzhan Huang, Jia Ren, Cecilia N. Arighi, Gang Li,
Catalina O. Tudor, Mengxi Lv, Jung-Youn Lee, Sheng-Chih Chen,
K. Vijay-Shanker, and Cathy H. Wu*

PART IV PROTEOMIC BIOINFORMATICS

- 17 Protein Identification from Tandem Mass Spectra by Database Searching 357
Nathan J. Edwards
- 18 Bioinformatics Analysis of Top-Down Mass Spectrometry Data
with ProSight Lite. 381
*Caroline J. DeHart, Ryan T. Fellers, Luca Fornelli, Neil L. Kelleher,
and Paul M. Thomas*
- 19 Mapping Biological Networks from Quantitative Data-Independent
Acquisition Mass Spectrometry: Data to Knowledge Pipelines 395
*Erin L. Crowgey, Andrea Matlock, Vidya Venkatraman,
Justyna Fert-Bober, and Jennifer E. Van Eyk*
- 20 Annotation of Alternatively Spliced Proteins and Transcripts
with Protein-Folding Algorithms and Isoform-Level Functional Networks. . . . 415
*Hongdong Li, Yang Zhang, Yuanfang Guan, Rajasree Menon,
and Gilbert S. Omenn*
- 21 Computational and Statistical Methods for High-Throughput
Mass Spectrometry-Based PTM Analysis 437
Veit Schwämmle and Marc Vaudel
- 22 Cross-Species PTM Mapping from Phosphoproteomic Data 459
Rima Chaudhuri and Jean Yee Hwa Yang
- Index* 471

Contributors

- CECILIA N. ARIGHI • *Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA*
- PEER BORK • *European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany; Max Delbrück Centre for Molecular Medicine, Berlin, Germany*
- MATTHEW P. CAMPBELL • *Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia*
- RIMA CHAUDHURI • *Charles Perkins Centre, School of Life and Environmental Sciences, University of Sydney, Camperdown, NSW, Australia*
- CHUMING CHEN • *Center for Bioinformatics and Computational Biology, Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA; Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA*
- JING CHEN • *UC San Diego, Center for Computational Biology and Bioinformatics, University of California, San Diego, La Jolla, CA, USA*
- SHENG-CHIH CHEN • *Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA*
- KAREN R. CHRISTIE • *The Jackson Laboratory, Bar Harbor, ME, USA*
- ERIN L. CROWGEY • *Nemours Alfred I. DuPont Hospital for Children, Wilmington, DE, USA*
- NATALIE L. DAWSON • *Institute of Structural and Molecular Biology, University College London, London, UK*
- CAROLINE J. DEHART • *Department of Chemistry and Molecular Biosciences, Proteomics Center of Excellence and the Robert H. Lurie Comprehensive Cancer Center at the Feinberg School of Medicine, Northwestern University, Evanston, IL, USA*
- HAYLEY DINGERDISSEN • *Department of Biochemistry and Molecular Medicine, George Washington University, Washington, DC, USA*
- HAROLD DRABKIN • *The Jackson Laboratory, Bar Harbor, ME, USA*
- NATHAN J. EDWARDS • *Department of Biochemistry, and Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- RYAN T. FELLERS • *Department of Chemistry and Molecular Biosciences, Proteomics Center of Excellence and the Robert H. Lurie Comprehensive Cancer Center at the Feinberg School of Medicine, Northwestern University, Evanston, IL, USA*
- JUSTYNA FERT-BOBER • *Advanced Clinical BioSystems Research Institute, Cedars Sinai Medical Center, Heart Institute, Los Angeles, CA, USA*
- LUCA FORNELLI • *Department of Chemistry and Molecular Biosciences, Proteomics Center of Excellence and the Robert H. Lurie Comprehensive Cancer Center at the Feinberg School of Medicine, Northwestern University, Evanston, IL, USA*
- ELISABETH GASTEIGER • *Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland*
- YUANFANG GUAN • *Department of Computational Medicine and Bioinformatics and the Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA*
- NAILA GULZAR • *Department of Biochemistry and Molecular Medicine, George Washington University, Washington, DC, USA*

- ATTILA GURSOY • *Center for Computational Biology and Bioinformatics, Koc University, Istanbul, Turkey; Computer Engineering, College of Engineering, Koc University, Istanbul, Turkey*
- ROBIN HAW • *Informatics and Biocomputing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada*
- HONGZHAN HUANG • *Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA*
- NEIL L. KELLEHER • *Department of Chemistry and Molecular Biosciences, Proteomics Center of Excellence and the Robert H. Lurie Comprehensive Cancer Center at the Feinberg School of Medicine, Northwestern University, Evanston, IL, USA*
- OZLEM KESKIN • *Chemical and Biological Engineering, College of Engineering, Koc University, Istanbul, Turkey; Center for Computational Biology and Bioinformatics, Koc University, Istanbul, Turkey*
- SU DATT LAM • *Institute of Structural and Molecular Biology, University College London, London, UK; School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia*
- JUNG-YOUN LEE • *Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA*
- JONATHAN G. LEES • *Institute of Structural and Molecular Biology, University College London, London, UK*
- GANG LI • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- HONGDONG LI • *Institute for Systems Biology, Seattle, WA, USA*
- QINGLIANG LI • *Department of Biochemistry, and Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- FREDERIQUE LISACEK • *Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland; Computer Science Department, University of Geneva, Geneva, Switzerland*
- MENGXI LV • *Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA*
- JULIEN MARIETHOZ • *Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland*
- MARIA J. MARTIN • *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
- ANDREA MATLOCK • *Advanced Clinical BioSystems Research Institute, Cedars Sinai Medical Center, Heart Institute, Los Angeles, CA, USA*
- RAJA MAZUMDER • *Department of Biochemistry and Molecular Medicine, George Washington University, Washington, DC, USA*
- RAJASREE MENON • *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA*
- PABLO MINGUEZ • *Department of Genetics and Genomics, Instituto de Investigacion Sanitaria, University Hospital Fundacion Jimenez Diaz (IIS-FJD), Madrid, Spain*
- BRION W. MURRAY • *Oncology Research Unit, Pfizer Worldwide Research and Development, San Diego, CA, USA*
- DARREN A. NATALE • *Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA*
- RUTH NUSSINOV • *Cancer and Inflammation Program, Leidos Biomedical Research, Inc., Frederick National Laboratory, National Cancer Institute, Frederick, MD, USA;*

- Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Sackler Institute of Molecular Medicine, Tel Aviv University, Tel Aviv, Israel*
- CLAIRE O'DONOVAN • *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
- GILBERT S. OMENN • *Departments of Computational Medicine and Bioinformatics, Internal Medicine, and Human Genetics, School of Public Health, University of Michigan, Ann Arbor, MI, USA*
- CHRISTINE A. ORENGO • *Institute of Structural and Molecular Biology, University College London, London, UK*
- NICOLLE H. PACKER • *Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia*
- ROBYN A. PETERSON • *Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia*
- RUDOLF T. PILLICH • *UC San Diego, Center for Computational Biology and Bioinformatics, University of California, San Diego, La Jolla, CA, USA*
- DEXTER PRATT • *UC San Diego, Center for Computational Biology and Bioinformatics, University of California, San Diego, La Jolla, CA, USA*
- SANGYA PUNDIR • *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
- JIA REN • *Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA*
- KAREN E. ROSS • *Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- VLADIMIR RYNKOV • *UC San Diego, Center for Computational Biology and Bioinformatics, University of California, San Diego, La Jolla, CA, USA*
- VEIT SCHWÄMMLE • *Department of Biochemistry and Molecular Biology, Protein Research Group, University of Southern Denmark, Odense, Denmark*
- SALIM SHAH • *Department of Biochemistry, and Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- K. VIJAY-SHANKER • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- IAN SILLITOE • *Institute of Structural and Molecular Biology, University College London, London, UK*
- ULRICH STELZL • *Otto-Warburg Laboratory, Max-Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany; Department of Pharmaceutical Chemistry, Institute of Pharmaceutical Sciences, University of Graz, Graz, Austria*
- PAUL M. THOMAS • *Department of Chemistry and Molecular Biosciences, Proteomics Center of Excellence and the Robert H. Lurie Comprehensive Cancer Center at the Feinberg School of Medicine, Northwestern University, Evanston, IL, USA*
- CATALINA O. TUDOR • *Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA*
- NURCAN TUNCBAG • *Graduate School of Informatics, Department of Health Informatics, Middle East Technical University, Ankara, Turkey*
- JENNIFER E. VAN EYK • *Advanced Clinical BioSystems Research Institute, Cedars Sinai Medical Center, Heart Institute, Los Angeles, CA, USA*
- MARC VAUDEL • *Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway*

- VIDYA VENKATRAMAN • *Advanced Clinical BioSystems Research Institute, Cedars Sinai Medical Center, Heart Institute, Los Angeles, CA, USA*
- ARUNACHALAM VINAYAGAM • *Department of Genetics, Harvard Medical School, Boston, MA, USA*
- QINGHUA WANG • *Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA; Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- REVATI WANI • *Oncology Research Unit, Pfizer Worldwide Research and Development, San Diego, CA, USA*
- DAVID WELKER • *UC San Diego, Center for Computational Biology and Bioinformatics, University of California, San Diego, La Jolla, CA, USA*
- JONATHAN WOODSMITH • *Otto-Warburg Laboratory, Max-Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany*
- CATHY H. WU • *Center for Bioinformatics and Computational Biology, Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA; Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA*
- GUANMING WU • *Informatics and Biocomputing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada; Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA*
- DONG XU • *Department of Computer Science and Bond Life Sciences Center, University of Missouri, Columbia, MO, USA*
- CHENG YAN • *Department of Biochemistry and Molecular Medicine, George Washington University, Washington, DC, USA*
- JEAN YEE HWA YANG • *School of Mathematics and Statistics, University of Sydney, Camperdown, NSW, Australia*
- QIUMING YAO • *Department of Computer Science and Bond Life Sciences Center, University of Missouri, Columbia, MO, USA*
- YANG ZHANG • *Department of Computational Medicine and Bioinformatics and the Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA*

Part I

Fundamentals of Protein Bioinformatics

Chapter 1

Protein Bioinformatics Databases and Resources

Chuming Chen, Hongzhan Huang, and Cathy H. Wu

Abstract

Many publicly available data repositories and resources have been developed to support protein-related information management, data-driven hypothesis generation, and biological knowledge discovery. To help researchers quickly find the appropriate protein-related informatics resources, we present a comprehensive review (with categorization and description) of major protein bioinformatics databases in this chapter. We also discuss the challenges and opportunities for developing next-generation protein bioinformatics databases and resources to support data integration and data analytics in the Big Data era.

Key words Bioinformatics, Database, Protein sequence, Protein structure, Protein family, Protein function, Protein mutation, Protein interaction, Pathway, Proteomics, PTM, Data integration, Data analytics, Big data

1 Introduction

Use of high-throughput technologies to study molecular biology systems in the past decades has revolutionized biological and biomedical research, allowing researchers to systematically study the genomes of organisms (Genomics) [1], the set of RNA molecules (Transcriptomics) [2], and the set of proteins including their structures and functions (Proteomics) [3]. Since proteins occupy a middle ground molecularly between gene and transcript and many higher levels of molecular and cellular structure and organization, and most physiological and pathological processes are manifested at the protein level, biological and biomedical scientists are increasingly interested in applying high-throughput proteomics techniques to achieve a better understanding of basic molecular biology and disease processes [4, 5].

The richness of proteomics data allows researchers to ask complex biological questions and gain new scientific insights. To support data-driven hypothesis generation and biological knowledge discovery, many protein-related bioinformatics databases, query facilities, and data analysis software tools have been developed

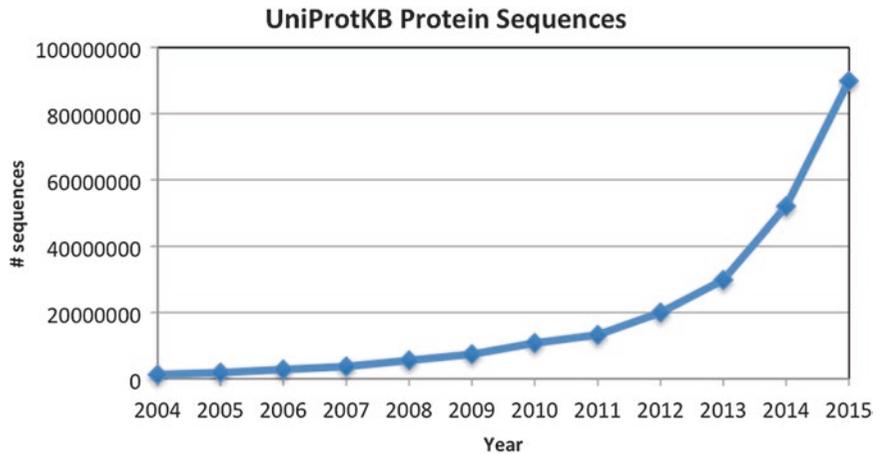


Fig. 1 The total number of protein sequences in UniProtKB. The diagram shows that as the result of the rapid development of genome sequencing projects, protein sequences archived in UniProtKB have increased dramatically in recent years

(http://www.oxfordjournals.org/our_journals/nar/database/cap/) to organize and provide biological annotations for proteins to support sequence, structural, functional, and evolutionary analyses in the context of pathway, network, and systems biology. With the recent extraordinary advances in genome sciences and Next-Generation Sequencing (NGS) technologies [6] that have uncovered rich genomic information in a huge number of organisms, new protein bioinformatics databases are also being introduced and many existing databases have been enhanced. As more and more genomes are sequenced, the protein sequences archived in databases have increased dramatically in recent years (*see* Fig. 1 for an example). This poses new challenges for computational biologists in building new infrastructure to support protein science research in the age of Big Data.

We present a summary review (with categorization and description) of protein bioinformatics databases and resources in Table 1. The databases and categories presented in Table 1 are selected from the databases listed in the Nucleic Acids Research (NAR) database issues and database collection, as well as the databases cross-referenced in the UniProtKB. The reason we choose them is because they: (1) are protein related and well grouped; (2) are well documented with papers and websites; (3) have been peer reviewed or/and selected by the UniProt consortium for UniProtKB database cross-references; and (4) are supposed to be well maintained.

Protein bioinformatics databases can be primarily classified as sequence databases, 2D gel databases, 3D structure databases, chemistry databases, enzyme and pathway databases, family and domain databases, gene expression databases, genome annotation

Table 1
Overview of protein bioinformatics databases

Category	DB short name	DB name	URLS	Ref.
Sequence databases	CCDS	The consensus CDS protein set database	https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi	[9]
	DDBJ	DNA Data Bank of Japan	http://www.ddbj.nig.ac.jp/	[10]
	ENA	European nucleotide archive	http://www.ebi.ac.uk/ena	[11]
	GenBank	GenBank nucleotide sequence database	https://www.ncbi.nlm.nih.gov/genbank/	[12]
	RefSeq ^a	NCBI reference sequence database	https://www.ncbi.nlm.nih.gov/refseq/	[13]
	UniGene	Database of computationally identified transcripts from the same locus	https://www.ncbi.nlm.nih.gov/unigene	[12]
	UniProtKB ^a	Universal Protein resource (UniProt)	http://www.uniprot.org/	[14]
2D gel databases	COMPLUYEAST-2DPAGE	2-DE database at Universidad Complutense de Madrid, Spain	http://compluyeast2dpage.dacya.ucm.es/	[15]
	REPRODUCTION-2DPAGE	2-DE database at Nanjing Medical University, China	http://reprod.njmu.edu.cn/cgi-bin/2d/2d.cgi	[16]
	SWISS-2DPAGE	2-DE database at Swiss Institute of Bioinformatics, Switzerland	http://world-2dpage.expasy.org/swiss-2dpage/	[17]
	World-2DPAGE ^a	The World-2DPAGE database	http://world-2dpage.expasy.org/repository/	[18]
3D structure databases	DisProt	Database of protein disorder	http://www.disprot.org/	[19]
	MobiDB	Database of intrinsically disordered and mobile proteins	http://mobidb.bio.unipd.it/	[20]
	ModBase	Database of comparative protein structure models	http://modbase.compbio.ucsf.edu/modbase.cgi/index.cgi	[21]
	PDBe ^a	Protein Data Bank at Europe	http://www.ebi.ac.uk/pdbe/	[22]
	PDBj ^a	Protein Data Bank at Japan	http://pdbj.org/	[23]
	PDBsum	Pictorial database of 3D structures in the Protein Data Bank	http://www.ebi.ac.uk/pdbsum/	[24]
	ProteinModelPortal	Protein model portal of the PSI-Nature structural biology knowledgebase	http://www.proteinmodelportal.org/	[25]
	RCSB-PDB ^a	Protein Data Bank at RCSB	http://www.pdb.org/	[26]
	SMR	Database of annotated 3D protein structure models	http://swissmodel.expasy.org/repository/	[27]
	Chemistry databases	BindingDB	The binding database	http://www.bindingdb.org/
ChEMBL ^a		Database of bioactive drug-like small molecules	https://www.ebi.ac.uk/chembl/	[29]
DrugBank		Drug and drug target database	http://www.drugbank.ca/	[30]

(continued)

**Table 1
(continued)**

Category	DB short name	DB name	URLs	Ref.	
Enzyme and pathway databases	MetaCyc/BioCyc ^a	MetaCyc database of metabolic pathways, BioCyc collection of pathway/genome databases	http://www.biocyc.org/	[31]	
	BRENDA ^a	BRaunschweig ENzyme DAtabase	http://www.brenda-enzymes.org	[32]	
	ENZYME	Enzyme nomenclature database	http://enzyme.expasy.org/	[33]	
	Reactome ^a	A knowledgebase of biological pathways and processes	http://www.reactome.org/	[34]	
	SABIO-RK	SABIO-RK: biochemical reaction kinetics database	http://sabiorck.h-its.org/	[35]	
	Signalink	A signalling pathway resource with multi-layered regulatory networks	http://signalink.org/	[36]	
	UniPathway	UniPathway: a resource for the exploration of metabolic pathways	http://www.unipathway.org	[37]	
	Family and domain databases	Gene3D	Structural and functional annotation of protein families	http://gene3d.biochem.ucl.ac.uk/Gene3D/	[38]
		HAMAP	High-quality automated and manual annotation of proteins	http://hamap.expasy.org/	[39]
		InterPro ^a	Integrated resource of protein families, domains and functional sites	http://www.ebi.ac.uk/interpro/	[40]
PANTHER		The PANTHER classification system	http://www.pantherdb.org/	[41]	
Pfam ^a		The Pfam protein families database	http://pfam.xfam.org/	[42]	
PIRSF ^a		A whole-protein classification database	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml	[43]	
PRINTS		Protein Motif fingerprint database	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/	[44]	
ProDom		Protein domain families database	http://prodrom.prabi.fr/prodom/current/html/home.php	[45]	
PROSITE ^a		Database of protein domains, families and functional sites	http://prosite.expasy.org/	[46]	
TIGRFAMs		ProtoNet	Automatic hierarchical classification of proteins	http://www.protonet.cs.huji.ac.il/	[47]
	SMART	Simple modular architecture research tool	http://smart.embl.de/	[48]	
	SUPEFAM	Superfamily database of structural and functional annotation	http://supfam.org	[49]	
	TIGRFAMs	TIGRFAMs protein family database	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	[50]	

Gene expression databases	<p>Bgee CleanEx Genevisible ExpressionAtlas^a</p>	<p>Database for gene expression evolution Database of gene expression profiles Search portal to normalized and curated expression data from Genevestigator Database of Differential and Baseline Expression</p>	<p>http://bgee.unil.ch http://cleanx.vital-it.ch/ http://genevisible.com/search http://www.ebi.ac.uk/gxa/home</p>	<p>[51] [52] [53] [54]</p>
Genome annotation databases	<p>Ensembl^a EnsemblBacteria EnsemblFungi EnsemblMetazoa EnsemblPlants EnsemblProtists Entrez Gene^a</p> <p>KEGG PATRIC UCSC^a VectorBase WBParaSite</p>	<p>Ensembl Eukaryotic genome annotation database Ensembl Bacteria genome annotation database Ensembl Fungi genome annotation database Ensembl Metazoa genome annotation database Ensembl Plants genome annotation database Ensembl Protists genome annotation database Database of Genes of Genomes in the Reference Sequence Collection Kyoto Encyclopedia of Genes and Genomes Bacterial Bioinformatics Resource Center UCSC Genome Bioinformatics Bioinformatics resource for invertebrate vectors of human pathogens WormBase ParaSite</p>	<p>http://www.ensembl.org/ http://bacteria.ensembl.org/ http://fungi.ensembl.org/ http://metazoa.ensembl.org/ http://plants.ensembl.org/ http://protists.ensembl.org/ http://www.ncbi.nlm.nih.gov/gene http://www.genome.jp/kegg/ http://patricbrc.org/ http://genome.ucsc.edu http://www.vectorbase.org/ http://parasite.wormbase.org</p>	<p>[55] [56] [56] [56] [56] [56] [57] [58] [59] [60] [61] [62]</p>
Organism specific databases	<p>ArachnoServer CGD ConoServer CTD dictyBase EchoBASE</p> <p>EcoGene euHCVdb EuPathDB FlyBase^a GenAtlas</p> <p>GeneCards GenoList</p> <p>Gramene H-InvDB HGNC</p> <p>HPA</p>	<p>ArachnoServer: Spider toxin database Candida Genome Database ConoServer: Cone snail toxin database Comparative Toxicogenomics Database Central resource for Dictyostelid genomics EchoBASE—an integrated post-genomic database for <i>E. coli</i>. Escherichia coli strain K12 genome database The European Hepatitis C Virus database Eukaryotic Pathogen Database Resources A Database of Drosophila Genes & Genomes A database on genes, functions and related diseases The Human Gene Database Integrated environment for the analysis of microbial genomes A comparative resource for plants H-Invitational Database HUGO Gene Nomenclature Committee Database The Human Protein Atlas</p>	<p>http://www.arachnoserver.org http://www.candidagenome.org/ http://www.conoserver.org/ http://ctdbase.org/ http://dictybase.org/ http://www.york.ac.uk/res/thomas/ http://www.ecogene.org/ https://euhcvdb.ibcp.fr/euHCVdb/ http://eupathdb.org/eupathdb/ http://flybase.org/ http://genatlas.medicine.univ-paris5.fr/ http://www.genecards.org/ http://genodb.pasteur.fr/cgi-bin/WebObjects/GenoList http://www.gramene.org/ http://www.h-invitational.jp/ http://www.genenames.org/ http://www.proteinatlas.org/</p>	<p>[63] [64] [65] [66] [67] [68] [69] [70] [71] [72] [73] [74] [75] [76] [77] [78] [79]</p>

(continued)

Table 1
(continued)

Category	DB short name	DB name	URLS	Ref.
	HUGE	A Database of Human Unidentified Gene-Encoded Large Proteins	http://www.kazusa.or.jp/huge/	[80]
	LegioList	Legionella pneumophila genome database	http://genolist.pasteur.fr/LegioList/	[81]
	Leprona	Mycobacterium leprae genome database	http://mycobrowser.epfl.ch/leprosy.html	[82]
	MaizeGDB	Maize Genetics and genomics Database	http://www.maizegdb.org/	[83]
	MGD ^a	Mouse Genome Database	http://www.informatics.jax.org/	[84]
	Micado	MICrobial Advanced Database Organization	http://genome.jouy.inra.fr/cgi-bin/micado/index.cgi	[85]
	OMIM	Online Mendelian Inheritance in Man	http://www.omim.org/	[86]
	neXtProt ^b	Exploring the universe of human proteins	http://www.nextprot.org/	[87]
	Orphanet	The portal for rare diseases and orphan drugs	http://www.orpha.net/consor/cgi-bin/home.php?Lang=GB	[88]
	PharmGKB	The Pharmacogenomics Knowledgebase	http://www.pharmgkb.org	[89]
	PomBase	The scientific resource for fission yeast	http://www.pombase.org/	[90]
	PseudoCAP	The Pseudomonas Genome Database	http://www.pseudomonas.com/	[91]
	RGD	Rat Genome Database	http://rgd.mcw.edu/	[92]
	Rouge	A Database of Rodent Unidentified Gene-Encoded Large Proteins	http://www.kazusa.or.jp/rouge/	[80]
	SGD	Saccharomyces Genome Database	http://www.yeastgenome.org/	[93]
	TAIR	The Arabidopsis Information Resource	http://www.arabidopsis.org/	[94]
	TuberculList	Mycobacterium tuberculosis strain H37Rv genome database	http://tuberculist.epfl.ch	[95]
	WormBase	C. elegans and related nematodes genetics and genomics database	http://www.wormbase.org/	[62]
	Xenbase	Xenopus laevis and tropicalis biology and genomics resource	http://www.xenbase.org/	[96]
	ZFIN	The Zebrafish Model Organism Database	http://zfin.org/	[97]
Phylogenomic databases	eggNOG	Database of orthologous groups and functional annotation	http://eggnog.embl.de/	[98]
	HOGENOM	Database of Homologous Genes from Fully Sequenced Organisms	http://pbil.univ-lyon1.fr/databases/hogenom/home.php	[99]
	HOVERGEN	Homologous Vertebrate Genes Database	http://pbil.univ-lyon1.fr/databases/hovergen.html	[100]
	InParanoid KO	Eukaryotic Ortholog Groups with inparalogs	http://inparanoid.sbc.su.se/	[101]
		Kyoto encyclopedia of genes and genomes orthology	http://www.genome.jp/kegg/	[102]
	OMA ^a	The OMA orthology database	http://omabrowser.org/	[103]
	OrthoDB	Database of Orthologous Groups	http://cegg.unige.ch/orthodb6	[104]
	PhylomeDB	Database for complete catalogs of gene phylogenies (phylores)	http://phylomedb.org/	[105]
	TreeFam	Database of animal gene trees	http://www.treefam.org	[106]

Polymorphism and mutation databases	BioMut	https://hive.biochemistry.gwu.edu/tools/biomuta/	[107]
	dbSNP ^a	https://www.ncbi.nlm.nih.gov/SNP/	[12]
	DMDM	http://bioinf.umbc.edu/dmdm/	[108]
Protein-protein interaction databases	BioGRID	http://thebiogrid.org	[109]
	DIP	http://dip.doe-mbi.ucla.edu/	[110]
	IntAct ^a	http://www.ebi.ac.uk/intact/	[111]
	MINT	http://mint.bio.uniroma2.it/mint/	[112]
	STRING	http://string-db.org	[113]
	Proteomic databases	MaxQB	http://maxqb.biochem.mpg.de/mxdb/
PaxDb		http://pax-db.org	[115]
PeptideAtlas ^a		http://www.peptideatlas.org	[116]
PRIDE ^a		http://www.ebi.ac.uk/pride	[117]
ProMEX		http://promex.pph.univie.ac.at/promex/	[118]
DEPOD ^a		http://www.koehnlab.de/depod/index.php	[119]
iPTMnet ^b		http://proteininformationresource.org/iPTMnet/	[120]
PTM databases	PhosPhAt ^a	http://phosphat.uni-hohenheim.de	[121]
	Phospho.ELM ^a	http://phospho.elm.eu.org	[122]
	PhosphoGrid ^a	http://www.phosphogrid.org	[123]
	PhosphoSitePlus ^a	http://www.phosphosite.org	[124]
	UniCarbKB ^a	http://www.umicarbkb.org/	[125]
	GO ^a	http://www.genontology.org/	[126]
	PRO	http://pir.georgetown.edu/pro/pro.shtml	[127]

(continued)

**Table 1
(continued)**

Category	DB short name	DB name	URLs	Ref.	
Specialized protein databases	Allergome	Allergome; platform for allergen knowledge	http://www.allergome.org/	[128]	
	CAZY	Carbohydrate-Active enZymes Database	http://www.cazy.org/	[129]	
	ESTHER	ESTERases and alpha/beta-Hydrolase Enzymes and Relatives database	http://bioweb.enscm.inra.fr/ESTHER/general?what=index	[130]	
	GPCRDB	Information system for G protein-coupled receptors (GPCRs)	http://www.gpcr.org/7tm/	[131]	
	IMGT	The International ImMunoGene Tics information system	http://www.imgt.org/	[132]	
	MEROPS ^a	MEROPS protease database	http://merops.sanger.ac.uk/	[133]	
	MoonProt	Moonlighting protein database	http://www.moonlightingproteins.org/	[134]	
	mycoCLAP	Characterized Lignocellulose-Active Proteins of Fungal Origin	https://mycoclap.fungalgenomics.ca/mycoCLAP/	[135]	
	PeroxiBase	The peroxidases database	http://peroxidase.toulouse.inra.fr/	[136]	
	REBASE	The Restriction Enzyme Database	http://rebase.neb.com/rebase/rebase.html	[137]	
	TCDB	Transporter Classification Database	http://www.tcdb.org/	[138]	
	Other (Miscellaneous) databases	ChiTaRS EvolutionaryTrace	Database of chimeric transcripts and rna-seq data Database of relative evolutionary importance of amino acids within a protein sequence	http://chitars.bioinfo.cnio.es/ http://mammoth.bcm.tmc.edu/ETserver.html	[139] [140]
		GeneWiki ^a	Wiki portal for the annotation of gene and protein function	http://en.wikipedia.org/wiki/Portal:Gene_Wiki	[141]
GenomeRNAi		Database of phenotypes from RNA interference screens in <i>Drosophila</i> and <i>Homo sapiens</i>	http://genomernai.dkfz.de/GenomeRNAi/	[142]	
PMAP-CutDB SOURCE		Proteolytic event database The Stanford Online Universal Resource for Clones and ESTs	http://www.proteolysis.org/ http://smd.princeton.edu/cgi-bin/source/sourceSearch	[143] [144]	

^aDatabases covered in the Subheading 3 of the chapter

databases, organism-specific databases, phylogenomic databases, polymorphism and mutation databases, protein-protein interaction databases, proteomic databases, PTM databases, ontologies, specialized protein databases, and other (miscellaneous) databases. Please visit <http://proteininformationresource.org/staff/chenc/MiMB/dbSummary2015.html> to access the databases reviewed in this chapter through their corresponding web addresses (URLs). For many of these databases, their identifiers can be mapped to UniProtKB protein AC/IDs [7]. Our coverage of protein bioinformatics databases in this chapter is by no means exhaustive. Our intention is to cover databases that are recent, high quality, publicly available, and are expected to be of interest to more users in the community. It is worth noting that certain databases can be classified into more than one category.

As an update to our previously contributed MiMB series chapter [8], we now focus on databases that are aligned with the content of this book and emphasize the types of data stored and related data access and data analysis supports. For each category of databases listed in Table 1, we select some representatives and describe them briefly in Subheading 2. In Subheading 3, we discuss the challenges and opportunities for developing next-generation protein bioinformatics databases and resources to support data integration and data analytics in Big Data era. We conclude the chapter in Subheading 4.

2 Databases and Resources Highlights

2.1 *Sequence Databases*

2.1.1 *RefSeq*

The National Center for Biotechnology Information Reference Sequence (NCBI RefSeq) database [13] provides curated non-redundant sequences of genomic regions, transcripts, and proteins for taxonomically diverse organisms including Archaea, Bacteria, Eukaryotes, and Viruses. RefSeq database is derived from the sequence data available in the redundant archival database GenBank [12]. RefSeq sequences include coding regions, conserved domains, variations etc. and enhanced annotations such as publications, names, symbols, aliases, Gene IDs, and database cross-references. The sequences and annotations are generated using a combined approach of collaboration, automated prediction, and manual curation [13]. The RefSeq release 73 on November 6, 2015 includes 54,766,170 proteins, 12,998,293 transcripts, and 55,966 organisms. The RefSeq records can be directly accessed from NCBI web sites by search of the Nucleotide or Protein databases, BLAST searches against selected databases, and FTP downloads. RefSeq records are also available through indirect links from other NCBI resources such as Gene, Genome, BioProject, dbSNP, ClinVar, Map Viewer, etc. In addition, RefSeq supports programmatic access through Entrez Programming Utilities [145].

2.1.2 UniProt

The UniProt Consortium consists of research teams from the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). The UniProt Consortium provides a central resource for protein sequences and functional annotations with four core database components to support protein bioinformatics research.

1. The UniProt Knowledgebase (UniProtKB) is the predominant data store for functional information on protein sequences with rich and accurate annotations (protein name or description, taxonomic information, classification, cross-reference and literature citation) [14]. The UniProtKB consists of two parts: UniProtKB/Swiss-Prot, which contains manually annotated records with information extracted from the literature and curator-evaluated computational analysis, and UniProtKB/TrEMBL, which contains computationally analyzed records with automatic annotation and classification. Comparative analysis and query for proteins are supported by UniProtKB extensive cross-references, functional and feature annotations, classification, and literature-based evidence attribution. The 2015_12 release on December 09, 2015 of UniProtKB/Swiss-Prot contains 550,116 sequence entries, comprising 196,219,159 amino acids, and 55,270,679 UniProtKB/TrEMBL sequence entries comprising 18,388,518,872 amino acids.
2. The UniProt Archive (UniParc) [146] is a comprehensive and non-redundant archival protein sequence database from all major publicly accessible resources. UniParc contains protein sequences and cross-references to their source databases. UniParc stores each unique protein sequence with a stable and unique identifier and tracks sequence changes in its source databases.
3. The UniProt Reference Clusters (UniRef) [147] are clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records. UniRef merges sequences and sub-fragments with 100 % (UniRef100), ≥ 90 % (UniRef90), or ≥ 50 % (UniRef50) identity and 80 % overlap with the longest sequences in the cluster (seed) into a single UniRef entry and select the highest ranked protein sequences as the cluster representatives.
4. The UniProt Proteomes [14] provides sets of proteins that are considered to be expressed by organisms whose genomes have been completely sequenced. A UniProt proteome consists of all UniProtKB/Swiss-Prot entries plus those UniProtKB/TrEMBL entries mapped to Ensembl Genomes for that proteome. Some well-studied model organisms and other organisms of interest to biomedical research and phylogeny have been manually and computationally [148] selected as reference proteomes.

The UniProt web site (<http://www.uniprot.org>) is the primary access point to its data and documentation. The site provides batch retrieval using UniProt identifiers; BLAST-based sequence similarity search; Clustal Omega-based sequence alignment; and Database identifier mapping [7]. The UniProt FTP download site provides batch download of protein sequence data in various formats, including flat file TEXT, XML, RDF and FASTA. Programmatic access to data and search result is supported via RESTful web services. For more details about UniProt databases, we refer the readers to Chapter 2 of this book.

2.2 2D Gel Databases: World-2DPAGE

The World-2DPAGE Constellation [18] is an effort of the Swiss Institute of Bioinformatics to promote and publish two-dimensional gel electrophoresis proteomics data online through the ExPASy proteomics server. The World-2DPAGE Constellation consists of three components:

1. **World-2DPAGE List** (<http://world-2dpage.expasy.org/list/>) contains references to known federated 2-D PAGE databases, as well as to 2-D PAGE-related servers and services.
2. **World-2DPAGE Portal** (<http://world-2dpage.expasy.org/portal/>) is a dynamic portal that serves as a single interface to query simultaneously worldwide gel-based proteomics databases that are built using the Make2D-DB package [149].
3. **World-2DPAGE Repository** (<http://world-2dpage.expasy.org/repository/>) is a public repository for gel-based proteomics data with protein identifications published in the literature. Mass-spectrometry-based proteomics data from related studies can also be submitted to the PRIDE database [117] so that interested readers can explore the data in the views of 2D-gel and/or MS.

The World-2DPAGE Constellation also provides a set of tools:

1. **Make2D-DB package** (ver. 3.10.2) is open source packages that can be used to build a user's own 2-D PAGE web site, access and integrate federated 2D-PAGE databases, portals, or data repositories.
2. **Melanie Viewer** (ver. 7.0) is a free viewer that can be used to visualize gels and related data obtained through the use of the full version of Melanie 2D electrophoresis gel analysis software.
3. **MIAPEGelDB** can be used to produce MIAPE-compliant gel experiments documents.

2.3 3D Structure Databases: wwPDB

The worldwide PDB (wwPDB, <http://www.wwpdb.org>) [150] was established in 2003 as an international collaboration to maintain a single and publicly available Protein Data Bank Archive

(PDB Archive) of macro-molecular structural data. The wwPDB member includes Protein Data Bank in Europe (PDBe) [22], Protein Data Bank Japan (PDBj) [23], Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) [26], and Biological Magnetic Resonance Bank (BMRB) [151]. The “PDB Archive” is a collection of flat files in three different formats: the legacy PDB format; the PDBx/mmCIF (<http://deposit.pdb.org/mmCIF/>) format; and the Protein Data Bank Markup Language (PDBML) [152] format. Each member site serves as a deposition, data processing, and distribution site for the PDB Archive, and each provides its own view of the primary data and a variety of tools and resources. As of December 1, 2015, there are 113,971 biological macromolecular structures in the wwPDB database including 37,049 distinct protein sequences, 30,099 structures of human sequences, 8,096 Nucleic Acid containing structures.

2.4 Chemistry Databases: ChEMBL

ChEMBL [29] is a large-scale bioactivity database containing binding, functional, in vivo absorption, distribution, metabolism, excretion, and toxicity (ADMET) information about drug-like bioactive compounds. ChEMBL data are manually curated from the published literature together with data drawn from other databases. ChEMBL data are standardized for using in many types of chemical biology and drug-discovery research problems. ChEMBL database can be accessed from a web-based interface where a variety of search and browsing functionality are provided. ChEMBL data are freely available from their FTP site in the formats of Oracle, MySQL, PostgreSQL, structure-data file (SDF), FASTA, and RDF. Programmatic access is also supported by a set of RESTful web services. The ChEMBL release 20 (prepared on Jan 14, 2015) contains 1,715,135 compound records, 1,463,270 compounds (of which 1,456,020 have mol files), 13,520,737 activities, 1,148,942 assays, 10,774 targets, and 59,610 documents.

2.5 Enzyme and Pathway Databases

2.5.1 MetaCyc and BioCyc

MetaCyc is a reference database of nonredundant, experimentally elucidated metabolic pathways and enzymes curated from the scientific literature [31]. MetaCyc stores pathways, compounds, proteins, protein complexes, and genes associated with these pathways with extensive links to protein sequence databases, nucleic acid sequence databases, protein structure databases, and literature. MetaCyc can also be used as a reference database to predict the metabolic network in sequenced genomes using Pathway Tools software [153] and machine-learning methods [154]. The 2015 release of MetaCyc includes 2,411 metabolic pathways, 13,074 reactions, 10,789 enzymes, 10,928 genes, 12,792 chemical compounds, 2,740 organisms, and 47,838 citations.

BioCyc is a collection of Pathway/Genome Databases (PGDBs) [31]. Each BioCyc PGDB contains the metabolic network of one organism predicted by the Pathway Tool software

using MetaCyc as a reference database. The BioCyc databases are organized into three tiers: Tier 1 databases are those that have received at least one person-year of literature-based curation. Tier 2 and Tier 3 databases are computationally predicted metabolic pathways. Web-based query, browsing, visualization, and comparative analysis tools are also provided from MetaCyc and BioCyc web sites. A collection of data files in different formats is provided for download. BioCyc also provides RESTful web services, MySQL server and Perl, Java, and Lisp APIs access to its data. The 2015 release of BioCyc includes 7,667 Pathway/Genome Databases.

2.5.2 BRENDA

BRENDA (BRAunschweig ENzyme DAtabase) [32] is an information system for functional and molecular properties of enzymes and enzyme-ligands obtained by manual extraction from literature, text and data mining, data integration, and computational predictions. BRENDA stores enzyme data in textual, single numeric, numeric range, and graphic formats. The content of BRENDA is based on the IUBMB (International Union of Biochemistry and Molecular Biology) enzyme classification system. BRENDA includes the following databases generated by a text mining approach:

1. **KENDA** contains kinetic values and kinetic expressions mined from PubMed abstracts.
2. **DRENDA** contains disease-related enzyme information (causal interaction, therapeutic application, diagnostic usage, and ongoing research) mined from PubMed abstracts using MeSH terms.
3. **FRENDA** contains references found in PubMed abstracts that have the enzyme name and organism combination.
4. **AMENDA** is a subset of FRENDA providing organism-specific information on the enzyme sources and the subcellular localization.

The user can access the data and information in BRENDA by searching (Quick Search, Advanced Search, Full text Search, Substructure Search, and Sequence Search) and browsing (TaxTree Explorer, EC Explorer, Ontology Explorer, and Genome Explorer). The search results can be downloaded as CSV file. The BRENDA release 2015.2 in July 2015 contains 6,759 enzymes.

2.5.3 Reactome

Reactome [34] is an open source, expert-curated, and peer-reviewed database of biological reactions and pathways with cross-references to major molecular databases. Reactome provides the visual representation of classical intermediary metabolism, signaling, innate and acquired immune function, transcriptional regulation, apoptosis and disease process, etc. The Reactome website supports the navigation of pathway knowledge and pathway-based

analysis and visualization of experimental or computational data. Interaction, reaction, and pathway data are downloadable as flat file, MySQL, BioPAX, SBML, and PSI-MITAB files. They are also accessible through RESTful web services. Software tools such as Pathway Browser, Analyze Data, Species Comparison, and Reactome FI Network are provided to support data mining and analysis of large-scale data sets. The Reactome release 54 in September 2015 contains 101,670 proteins, 74,357 complexes, 68,659 reactions, and 20,261 pathways.

2.6 Family and Domain Databases

2.6.1 InterPro

InterPro [40] is an integrated resource of predictive models or “signatures” representing protein domains, families, regions, repeats, and sites from major protein signature databases including CATH-Gene3D [38], HAMAP [37], PANTHER [41], Pfam[42], PIRSF [43], PRINTS [44], ProDom [45], PROSITE [46], SMART [48], SUPERFAMILY [49], and TIGRFAMs [50]. Each entry in the InterPro database is annotated with a descriptive abstract name and cross-references to the original data sources, as well as to specialized functional databases. The search by sequence or domain architecture is provided by the InterPro web site. The InterPro signatures in XML format are available via anonymous FTP download. InterPro also provides a software package InterProScan [155] that can be used locally to scan protein sequences against InterPro’s signatures. Programmatic access to InterProScan is possible via RESTful and SOAP web service APIs. The InterPro BioMart [156] allows users to retrieve InterPro data from a query-optimized data warehouse that is synchronized with the main InterPro database, and to build simple or complex queries and control the query results through a unified interface. The InterPro release 54.0 on October 15, 2015 includes 28,462 entries containing signatures of 19,110 families, 8,191 domains, 284 repeats, 115 active sites, 74 binding sites, 672 conserved sites, and 16 PTMs.

2.6.2 Pfam

Pfam is a database of protein families represented as multiple sequence alignments and Hidden Markov Models (HMMs) [42]. Pfam entries can be classified as Family (related protein regions), Domain (protein structural unit), Repeat (multiple short protein structural units), or Motifs (short protein structural unit outside global domains). Related Pfam entries are grouped into clans based on sequence, structure, or profile-HMM similarity. The Pfam database web site provides a search interface for querying by sequence, keyword, domain architecture, or taxonomy, and browse interfaces for analyzing protein sequences for Pfam matches and viewing Pfam annotations in domain architectures, sequence alignments, interactions, species, and protein structures in PDB[26]. The Pfam data can be downloaded from its FTP site or programmatically accessed through RESTful web service APIs. The Pfam release 28.0 in May 2015 contains 16,230 families.

2.6.3 PIRSF

The PIRSF classification system [43] provides comprehensive and non-overlapping clustering of UniProtKB [14] sequences into a hierarchical order to reflect their evolutionary relationships based on whole proteins rather than on the component domains. The PIRSF system classifies the protein sequences into families, whose members are both homologous (evolved from a common ancestor) and homeomorphic (sharing full-length sequence similarity and a common domain architecture) [43]. The PIRSF family classification results are expert-curated based on literature review and integrative sequence and functional analysis. The classification report shows the information on PIRSF members and general statistics, family and function/structure relationships, database cross-references, and graphical display of domain and motif architecture of seed members or all members. The web-based PIRSF system has been demonstrated as a useful tool for studying the function and evolution of protein families [43]. It provides batch retrieval of entries from the PIRSF database. The PIRSF scan allows searching a query sequence against the set of fully curated PIRSF families with benchmarked Hidden Markov models. The PIRSF membership hierarchy data is also available for FTP download. The current release of PIRSF contains 11,800 families, which cover 5,407,000 UniProtKB protein sequences.

2.6.4 PROSITE

PROSITE [46] is a database of documentation entries describing protein domains, families, and functional sites as well as associated patterns and profiles to identify them. The entries are derived from multiple alignments of homologous sequences and have the advantage of identifying distant relationships between sequences. PROSITE includes a collection of ProRules based on profiles and patterns of functionally and/or structurally critical amino acids that can be used to increase PROSITE's discriminatory power [46]. The PROSITE web site provides keyword-based search and allows browsing by documentation entry, ProRule description, taxonomic scope, and number of positive hits. The software tool ScanProsite [157] supports three options for users to scan proteins for matches to PROSITE motifs or their own sequence patterns: (1) scan protein sequence against the PROSITE motifs; (2) scan motifs against a protein sequence database; (3) submit protein sequences and motifs and scan them against each other. The PROSITE documentation entries and related tools can be downloaded from its FTP site. The PROSITE release 20.120 on November 4, 2015 contains 1,742 documentation entries, 1,309 patterns, 1,139 profiles, and 1,138 ProRules.

2.7 Gene Expression Databases: Expression Atlas

The Expression Atlas database [54] provides gene, protein, and splice variant expression patterns in different cell types, organism parts, biological and experimental conditions. The high quality Microarray and RNA-Seq data imported from ArrayExpress[158]

and Gene Expression Omnibus [12] were manually curated, annotated, and processed using standardized analysis methods to detect the expression patterns under the original experimental conditions. Expression Atlas consists of two components: Baseline Atlas and Differential Atlas. The Baseline Atlas is about genes and their expression pattern under “normal” conditions using only RNA-Seq data. The Differential Atlas is about genes that are up- or down-regulated in differential biological or experimental conditions using both Microarray and RNA-Seq data. Expression Atlas web interface supports queries of both the Baseline Atlas and Differential Atlas by gene, protein, and splice variant. Searches for sample attributes and experimental conditions are also supported. All Expression Atlas analysis results can be downloaded from their FTP site. The differential expression data and meta-data can be used in the R Bioconductor (<https://www.bioconductor.org/>) package. The APIs to programmatically access Expression Atlas are under development. The October 29, 2015 release of Expression Atlas contains 2,373 datasets (93,057 assays).

2.8 Genome Annotation Databases

2.8.1 Ensembl

Ensembl is a genome annotation database that provides up-to-date annotations for chordates and model organism genomes [55]. Additional metazoan genomes are available from EnsemblMetazoa [56], plant and fungal genomes are available from EnsemblPlants [56] and EnsemblFungi [56], unicellular eukaryotic and prokaryotic genomes are available from EnsemblProtists [56] and EnsemblBacteria [56]. Ensembl supports a variety of access routes to their data. Small data sets can be exported from online search results. Large datasets or complex analyses can be accessed from MySQL server, Perl, and RESTful APIs. Complex cross databases queries are supported by the BioMart data mining tool [156]. The whole database can be downloaded from an FTP site in FASTA, EMBL, GenBank, GVF, VCF, VEP, GFF formats or through MySQL dumps. In addition, Ensembl also provides a set of data processing software tools, for example, Variant Effect Predictor, BLAST/BLAT, Assembly converter, ID History converter, etc. The Ensembl release v83 in September 2015 contains 69 species with annotations for gene and transcript, gene sequence evolution, genome evolution, sequence and structural variants, and regulatory elements.

2.8.2 Entrez Gene

Entrez Gene [57] is a NCBI gene-specific database that provides GeneIDs (unique integer identifiers) for genomes that have been completely sequenced. The data in Entrez Gene database (nomenclature, map location, gene products and attributes, markers, phenotypes, citations, sequences, variations, maps, expression, homologs, protein domains, etc.) are results of manual curation and automated computational analysis of data from RefSeq [13]

and many other NCBI databases [12]. The data in Entrez Gene database can be accessed in several ways: (1) query Entrez from the NCBI home page and display the results in Gene, (2) enter a query in any Entrez query bar and restrict the database search to Gene, (3) cross links from other NCBI resources such as GenBank, BLAST, RefSeq, or Map Viewer. Entrez Gene data can be downloaded from the NCBI FTP site and accessed by Entrez Programming Utilities [145]. The Entrez Gene release on December 4, 2015 includes 13,778 taxa and 12,841,400 genes.

2.8.3 UCSC

UCSC Genome Browser database [60] contains large collections of genome assemblies and annotations for vertebrate and selected model organisms. The major sources of genome annotations include RefSeq, GENCODE, Ensembl, GenBank, ENCODE, RepeatMasker, dbSNP, the 1000 Genome project and other resources. In addition to Genome Browser, the UCSC bioinformatics group also provides web-based and command-line-based tools to facilitate the use of genome annotation data. For example, BLAT can be used to quickly find sequences of 95 % and greater similarity and 25 bases or more in length. The Table Browser can retrieve the data associated with a track in Genome Browser and calculate intersections between tracks. The Variant Annotation Integrator can associate UCSC Genome Browser annotations with the user-uploaded variants. The Gene Sorter can be used to show expression, homology, and other information on groups of genes. User data can be viewed together with UCSC annotations via “custom track,” “track data hubs,” “assembly hub,” and “Genome Browser in a Box (GBiB)” [159]. Genome data and source codes are downloadable. UCSC Genome Bioinformatics group also provides public MySQL server access. Currently (December 11, 2015), there are 95 genomes in the UCSC Genome Browser database.

2.9 Organism Specific Databases

2.9.1 FlyBase

FlyBase [72] is a database of *Drosophila melanogaster* related genetic and genomic information. The sequence and annotation data for *Drosophila melanogaster* genome assembly can be downloaded from the FlyBase FTP site in multiple formats (GFF3, FASTA, GTF, Chado XML, and Chado PostgreSQL dump). FlyBase uses generic genome browser 2 (GBrowse 2) to display the genome annotations and genome-aligned evidence on the reference genome assembly. FlyBase database can be searched for genes, alleles, aberrations, and other genetic objects, phenotypes, sequences, stocks, images and movies, and controlled terms. FlyBase provides a standalone BLAST server for 50 different arthropod genomes and supports query results analysis such as hit list refinement and batch download. The latest FlyBase is FB2015_05 released on November 20, 2015 that consists of 212,991 references, 141,104 stocks, and 1,258 images.

2.9.2 MGD

The Mouse Genome Database (MGD) [84] is a database of integrated genomic, genetic, and biological data on the laboratory mouse that is a model for translational research. MGD integrates mouse genome annotations from NCBI, Ensembl, and Havana into a single non-redundant resource. MGD is the authoritative source for the unified catalog of mouse genome features, Gene Ontology (GO) annotations (functional associations) of mouse protein-coding genes, and mouse phenotype annotations. The Human-Mouse: Disease Connection (<http://www.diseasemodel.org>) is a translational research tool that provides simultaneous access to human-mouse genomic, phenotypic, and genetic disease information. MGD uses a powerful new genome browser called JBrowse [160] to integrate mouse gene and protein annotations with large-scale sequence data. In addition to online search tools for genes, genome features and maps, phenotypes, alleles and disease models, gene expression, GO functional annotations, strains, SNPs and polymorphisms, sequences, references, and vocabularies, MGD also provides bulk data download as FTP reports and batch query tool and programmatic access by Web services and BioMart [156]. MGD is updated on a weekly basis.

2.9.3 neXtProt

neXtProt [87] is a new protein-centric knowledge platform and serves as a central hub for all knowledge about human proteins. neXtProt integrates high-quality and manually curated UniProt/Swiss-Prot entries with large amount of additional human protein-related information from other resources such as Human Protein Atlas [79], ArrayExpress [158], UniGene [12], PeptideAtlas [116], Gene Ontology Annotation [126], Ensembl [55], dbSNP [12], etc. Ontologies and controlled vocabularies (CVs) are extensively used in neXtProt to support consistent annotation and data retrieval. neXtProt's Google-like search interface supports free text search and complex queries with results displayed as lists or short summaries. neXtProt provides export functionality for protein entries in TEXT, Excel, FASTA, and XML formats and bulk download from the FTP site. neXtProt release on September 1, 2015 contains 20,066 protein entries, 153,556 controlled vocabularies, and 465,706 publications.

2.10 Phylogenomic Databases: OMA

The Orthologous Matrix (OMA) [103] is a method and associated database that infers evolutionary relationships among complete proteomes. OMA's inference algorithm includes three steps: (1) infer homologous sequences (sequences of common ancestry); (2) infer orthologous pairs (subsets of homologs related by speciation events); (3) cluster orthologs into: (a) OMA groups (cliques of orthologous pairs) and (b) HOGs (groups of genes descended from a common ancestral gene in a given taxonomic range). OMA can be accessed through the OMA browser and programmatic interfaces. OMA genomes including all-against-all computations

can be downloaded with an OMA stand-alone program to do orthology prediction using the user's custom data. The OMA release in September 2015 contains 1,970 species, 1,001,242 OMA groups, and 10,129,468 proteins.

2.11 Polymorphism and Mutation **Databases: dbSNP**

The NCBI dbSNP database [12] is a database for short genetic variations from a variety of organisms. dbSNP catalogs single nucleotide variations, short nucleotide insertions and deletions, short tandem repeats, and microsatellites. The dbSNP homepage provides a search interface for querying variations by simple term or complex queries. The details of matched variation records are displayed as the Reference SNP Cluster Report that contains a summary of the allele, mapping information in Human Genome Variation Society (HGVS) nomenclature, gene-centric view, map table with chromosomal coordinates, variation view, and link to the 1000 Genomes Browser. dbSNP integrates disease-related variations collected by OMIM [86]. dbSNP variation data are accessible through links from other NCBI databases. dbSNP data can also be downloaded from a FTP site and accessed by EUtils API (<https://www.ncbi.nlm.nih.gov/books/NBK25500/>). dbSNP build 146 on November 24, 2015 for Homo sapiens contains 150,482,731 RefSNP Clusters; among them 100,135,281 are validated.

2.12 Protein-Protein Interaction **Databases: IntAct**

IntAct [111] is an open source database and toolkit for the storage, presentation, and analysis of rich curated molecular interaction data in community accepted standard formats. IntAct provides relevant experimental details of protein interactions curated from literature or directly deposited. All the entries in the database are fully compliant with the IMEx [162] guidelines and MIMIx [163] standard. The IntAct web site provides multiple search functionalities: (1) search by anything that might be related to interactions, for example, gene name, identifiers, GO term, publication, and experimental method etc.; (2) search on four ontologies: Gene Ontology [126], InterPro [40], PSI-MI [164], ChEBI [165]; (3) draw all or part of a chemical structure and search for chemical compounds. IntAct data is released monthly and available as FTP download. IntAct release 194 on December 2, 2015 consists of 577,297 binary interactions from 13,952 curated publications and 1,378 biological complexes.

2.13 Proteomics Databases

2.13.1 PeptideAtlas

PeptideAtlas [116] provides an approach and framework to archive proteomic data that enables the data exchange and integration with genomic data. PeptideAtlas statistically validates peptides identified by high-throughput tandem mass spectrometry (MS/MS) experiments and maps peptide sequences to eukaryotic genomes. PeptideAtlas uses a uniform statistical validation process to ensure consistent and high-quality peptide and protein identifications. The raw data used to build PeptideAtlas includes raw

MS/MS files, MS/MS files in mzXML[166] format, and SEQUEST [167] search results. The user can also download PeptideProphet [168] results and ProteinProphet [169] outputs. The PeptideAtlas builds are available for download or browsing via the PeptideAtlas web interface. As of December 7, 2015, there are in total 72 builds covering 19 organisms.

2.13.2 PRIDE

The PRoteomics IDentifications database (PRIDE) [117] is a repository for mass-spectrometry based proteomics data including identifications of proteins, peptides, and post-translational modifications that have been described in the scientific literature, together with supporting mass spectra and related technical and biological metadata. PRIDE supports tandem MS (MS/MS) and Peptide Fingerprinting datasets with search/analysis workflows originally analyzed by the submitters. PRIDE provides several services such as the Protein Identifier Cross-Reference (PICR) [170], the Ontology Lookup Service (OLS) [171], and Database on Demand [172]. The data in PRIDE database can be accessed in different ways: (1) The PRIDE web interface can be used to explore all public datasets currently available in the repository; (2) Batch data retrieval and integration with other databases can be achieved by PRIDE BioMart [156]; (3) PRIDE public experiments data in mzData (<http://www.psidev.info/mzdata>) and PRIDE XML formats can be downloaded via FTP, Aspera, and HTTP; (4) A set of RESTful web services can be used to get programmatic access to data in the PRIDE repository. PRIDE supports submissions of protein and peptide identification/quantification data with the accompanying mass spectral evidence by following the ProteomeXchange (PX) consortium [173] guidelines. PRIDE also provides a set of software tools: PRIDE Converter 2 for converting common mass spectrometry data formats into PRIDE XML for data submission, and PRIDE Inspector for visualizing and analyzing MS dataset, such as mzML [174], mzIdentML (<http://www.psidev.info/mzidentml>), and PRIDE XML. As of December 8, 2015, PRIDE repository includes 3,774 projects and 55,873 assays.

2.14 PTM Databases

2.14.1 DEPOD

The human DEPhosphorylation Database (DEPOD) [119] is a comprehensive, high-quality, manually curated database for human phosphatases, their experimentally verified protein and non-protein substrates, dephosphorylation sites, involved pathways with cross-references to kinases, and small molecule modulators. The human phosphatase substrate information is integrated from a variety of sources including “dephosphorylation” post-translational modification data in Human Protein Reference Database [175], “dephosphorylation” interaction data from PSICQUIC service [176], substrate information from UniProt annotation [14], and scientific literature from PubMed and Google. DEPOD database can be browsed by human phosphatases, protein substrates, non-protein

substrates, pathways, and phosphatase-substrate networks. DEPOD also allows direct deposit of substrate candidates for human active phosphatases. The human active phosphatase data can be downloaded in XSLX format. The human phosphatase-substrate interaction and dephosphorylation sites data are available for download in PSI-MI Tab 2.5 format. In addition, phosphatases and substrates mapped onto KEGG [58], NCI Nature PID and Reactome [34] pathways are available in TXT format. The latest release of DEPOD on August 15, 2015 contains 228 human active and 11 inactive phosphatases (194 phosphatases have substrate), 298 protein substrates, 89 nonprotein substrates, 1,096 dephosphorylation interactions, 213 KEGG pathways, 206 NCI Nature PID pathways, and 560 Reactome pathways.

2.14.2 *iPTMnet*

iPTMnet [120] is an integrated resource for protein post-translational modification network discovery that combines text mining, data mining, and ontological representation to capture rich PTM information, including PTM enzyme-substrate relationships, PTM-specific protein-protein interactions (PPIs), and PTM conservation across species to support PTM analysis in the context of systems biology. It employs the RLIMS-P [177] and eFIP [178] text mining tools developed by the PIR group for full-scale mining of PubMed abstracts to identify PTM information (kinase, substrate, and site) and phosphorylation-dependent PPIs. Experimentally observed PTMs, including high-throughput proteomic data from curated PTM databases, are incorporated. Proteins and PTM protein forms (proteoforms) are organized using the Protein Ontology (PRO) [127], enabling representation and annotation of forms modified on combinations of PTM sites and orthologous relationships between forms. iPTMnet thus serves as an integrated resource that connects knowledge about biologically relevant modified proteins from disparate sources. Covering seven major PTM types (phosphorylation, acetylation, ubiquitination, methylation, glycosylation, sumoylation, and myristoylation), the current iPTMnet database contains more than 250,000 PTM sites in more than 45,000 modified proteins, along with more than 1,000 PTM enzymes for human, mouse, rat, yeast, Arabidopsis, and several other organisms. The web portal supports online search and visual analysis for scientific queries. For more details about iPTMnet database, we refer the readers to Chapter 16 of this book.

2.14.3 *PhosPhAt*

The Arabidopsis Protein Phosphorylation Site Database (PhosPhAt) [121] catalogs published information on large-scale mass spectrometry experiments that have identified phosphorylation sites in Arabidopsis. It contains information about the peptides, their annotated biological functions, and experimental and analytical contexts as well as information about kinase-substrate relationships manually

curated from the literature. In addition, PhosPhAt provides a plant-specific phosphorylation site predictor trained using serine, threonine, and tyrosine phosphorylation (pSer, pThr, pTyr) experimental data. The user can access the precomputed prediction using *Arabidopsis* gene identifiers or do “on-the-fly” prediction of phosphorylation of user-submitted protein sequences. Both the experimentally determined phosphorylation sites and high confidence predicted sites are available for download. As of December 8, 2015, PhosPhAt includes 9,159 experimental phosphoproteins with 19,100 unique tryptic phosphopeptides, and 31,916 predicted proteins with 2,176,360 predicted phosphosites.

2.14.4 *Phospho.ELM*

Phospho.ELM [122] is a manually curated database of experimentally verified eukaryotic phosphorylation sites. Each entry in the Phospho.ELM database is manually annotated with information about the phosphorylated proteins, the positions of known phosphorylations, the kinases responsible for phosphorylation, and literature citations. Additional information such as structure, interaction partners, subcellular compartment, and tissue specificities is also provided whenever they are available. Phospho.ELM data can be searched from its web interface. The data sets are also available for download upon request. PhosphoBlast server can be used to search proteins (UniProt ID/AC or amino acid sequence) against the curated dataset of phosphorylated peptides. Phospho.ELM (v9.0, September 2010) contains 8,718 substrate proteins covering 3,370 tyrosine, 31,754 serine, and 7,449 threonine instances.

2.14.5 *PhosphoGrid*

PhosphoGrid [123] is a database of experimentally verified in vivo protein phosphorylation sites of *Saccharomyces cerevisiae* curated from the literature. Both high-throughput MS phosphoproteomics studies and focused low-throughput analyses of individual proteins or complexes are integrated into PhosphoGrid. Each in vivo phosphorylation site is annotated by a hierarchy of experimental evidence codes, experimentally defined protein kinases and/or phosphatases, specific condition(s) under which the phosphorylation event occurs and the effect(s) of phosphorylation on protein function. The user can search PhosphoGrid web-based interface for any substrate, protein kinase, or phosphatase. Each record is cross-referenced with BioGRID [109], *Saccharomyces* Genome Database (SGD) [93], NCBI protein database [12], and its original PubMed articles. The latest release of PhosphoGrid contains 20,177 phosphorylation sites, 3,011 kinases, 266 phosphatases, and 563 publications.

2.14.6 *PhosphoSitePlus*

PhosphoSitePlus (PSP) [124] is a curated and highly interactive systems biology knowledgebase for studying experimentally observed mammalian post-translational modifications (PTMs) and

their roles in the regulation of biological process. PSP provides a comprehensive coverage of protein phosphorylation, acetylation, methylation, ubiquitination, and O-glycosylation. PSP includes structural and functional information about the topology, biological function, and regulatory significance of modification sites integrated from both low- and high-throughput (LTP and HTP) data. The homepage of PSP includes “Simple Search” that allows query of all known phosphorylation sites in a specific protein and “Advanced Search” that allows search by protein, sequence, or reference. PSP also supports retrieval of a list of modified sites that possess certain specified attributes and browsing curated MS/MS records by disease type, cell line, and tissue. Multiple types of datasets and tools are available for download such as PTMVar datasets, modification site datasets, regulatory sites, disease-associated sites, kinase-substrate datasets, Cytoscape plugin, etc. The latest release of PSP (accessed on December 9, 2015) contains 52,872 proteins, 21,619 low-throughput (LTP) sites, 456,434 high-throughput (HTP) MS sites, 2,130,888 MS peptides, and 19,704 curator-reviewed papers.

2.14.7 UniCarbKB

UniCarbKB [125] is a curated knowledgebase for glycomics and glycobiology research. UniCarbKB provides comprehensive information about the structures, pathways, and networks involved in glycosylation and glycol-mediated processes. UniCarbKB integrates GlycoSuiteDB [179] and EUROCarbDB [180] to provide a unified portal to support glycol-bioinformatics research and knowledge dissemination. The content of UniCarbKB is mainly eukaryotic glycoproteins curated from GlycoSuiteDB and a selected few datasets from EUROCarbDB. The data in GlycosuiteDB, EUROCarbDB, and GlycoBase [181] can be queried by taxonomy, tissue, protein name, protein accession, and composition. Glycan structures can be searched using carbohydrate sequences in GlycoCT format. The user can browse the curated collection of proteins or search them by name. Glycan Builder provides a GUI interface for building and displaying glycan structures. GlycoDigest is a tool that simulates exoglycosidase digestion, based on controlled rules acquired from expert knowledge and experimental evidence available in GlycoBase. The latest release of UniCarbKB (accessed on December 9, 2015) contains 899 Glycoproteins, 3,238 GlycoSuite structures, 520 UniCarb-DB MS/MS datasets, and 909 publications.

2.15 *Ontology Databases: Gene Ontology (GO)*

The Gene Ontology (GO) [126] is a bioinformatics effort to create the consistent computational representation of gene functions at the molecular, cellular, and tissue system levels across all organisms. GO provides a controlled vocabulary of terms (ontologies) to describe gene products in terms of their biological processes, cel-

lular components, and associated molecular functions. The use of GO terms enables uniform queries and association across many biological databases. From the GO web site, the user can search for GO terms, annotations to gene products, and metadata across multiple species and perform GO enrichment analysis. The GO web site supports the download of the gene association files (Annotation), Gene Ontology (Ontology), and mappings of GO terms to those in a number of external vocabularies (Mapping). The Gene Ontology as of December 8, 2015 consists of 29,033 biological process terms, 4,039 cellular component terms, and 10,920 molecular function terms.

**2.16 Specialized
Protein Databases:
MEROPS**

MEROPS [133] is an integrated database of information about peptidases (also termed proteases, proteinases, and proteolytic enzymes) and the proteins that inhibit them. A homologous set of peptidases and protein inhibitors are grouped into peptidase and inhibitor species. Species are grouped into families that contain statistically significant similarities in amino acid sequence. Families are grouped into clans that contain related structures. Both family (subfamily) and clan can be browsed by index page with links to their summary page. Each peptidase has a summary page that can be browsed by name, identifier, gene name, organism, and substrates. The peptidase summary page includes information on gene structure, alignment, tree, sequences and their features, distribution, structure, literature, human EST, mouse EST, substrates, inhibitors, and pharmacological modulators. The MEROPS database can be searched for peptidases or inhibitors, peptidases or inhibitor genes, or structures of peptidases or inhibitors. Users can also search via specificity, organism, and citation. MEROPS supports searching peptidase and protein inhibitor sequences with a protein or nucleotide query sequence by WU-BLAST. MEROPS also provides batch substrate cleavage analysis. MEROPS allows online submission of protein cleavage sites; however login is required for data download.

**2.17 Other
(Miscellaneous)
Databases: Gene Wiki**

Gene Wiki [141] is a collection of community-written Wikipedia articles about human genes in the NCBI Gene database [57]. Gene Wiki starts with a set of seed stub Wikipedia articles, populated and expanded by community contributors with focus on the functions and disease relevance of the gene and corresponding protein. Gene Wiki has an automated system to keep the article structures in sync with the data from trusted primary databases and uses the WikiTrust [161] reputation system to assess and display the trustworthiness of authors and their contributions. Gene Wiki has over 10,000 distinct gene pages, spanning 2.07 million words and 82 megabytes of data.

3 Challenges and Opportunities

Although a large number of protein bioinformatics databases and resources have been developed to catalog and store different information about proteins, there are challenges and opportunities in developing next-generation databases and resources to facilitate data integration, data-driven hypothesis generation, and biological knowledge discovery. Recent rapid developments in high-throughput sequencing technologies bring molecular biology researchers to the age of Big Data, where the research paradigm has shifted from hypothesis-driven to data-driven. Big Data opens new avenues to study molecular biology as well as brings new challenges for computational biologists to explore ways to efficiently manage and analyze data, and eventually turn data into usable and actionable knowledge. Next, we will review and discuss some recent technology developments that can help in addressing some of the challenges.

3.1 *Characteristics of Big Data*

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate (Wikipedia, <https://www.wikipedia.org/>). More specifically, Big Data has the following characteristics:

1. **Volume** The size of data is definitely an important aspect of Big Data. Large volumes of data demand scalable storage solutions and distributed information processing and retrieval.
2. **Variety** The types of data determine how the data will be analyzed. The heterogeneity of data requires non-trivial analysis methods.
3. **Velocity** The speed with which the data are generated and processed challenges novel real-time data analytics.
4. **Variability** The inconsistency of data calls for effective data management and handling.
5. **Veracity** The accuracy of data analysis depends on high-quality data and data capture methodology.

3.2 *Data Storage and Management*

The first challenge computational biologists have to face is the efficient storage and management of large volumes of data. In addition to better hardware support, massive parallel storage systems (distributed file systems, cluster file systems, and parallel file systems) have been explored. Examples include the Lustre [182] and Hadoop Distributed File System (HDFS) [183]. On top of that we need frameworks for user-specific solutions where several tools have been developed. Apache Hive [184] is a distributed data warehouse framework for analyzing data stored in HDFS and compatible systems using a SQL-like language called HiveQL. Apache Pig [185] further simplifies complex data analysis using simpler

scripting language targeting domain experts. Traditional relational database management systems often have difficulty handling Big Data because they lack horizontal scalability, require hard consistency, and become very complex when dealing with large volume of heterogeneous data. Non-relational databases (NoSQL) are alternative to Big Data storage and management because they focus on scalability and flexibility. The popular NoSQL database management systems include key-value stores, columnar databases, graph databases, and document-oriented databases.

3.3 Data Analytics

Data storage and management is only one side of the coin. In the field of biomedical research and healthcare systems, the purpose of high-throughput omics studies is to turn biomedical data into knowledge. In order to accomplish the goal of personalized medicine and better treatments, we need scalable computational facilities and efficient data analytics frameworks. Compared to traditional HPC cluster computing, cloud computing emerges as an economical solution to large-scale data analysis. Hosting large-volume high-throughput data in the cloud is changing the way the analysis is done. Instead of moving data to the analysis code, code is now moving to the data. In addition, novel and efficient machine learning and data mining algorithms and computational frameworks are also essential to the success of turning data into knowledge. Apache Spark [186] is a recently developed fast and general computing engine for large-scale lightning-fast in-memory clustering computing. It supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming for scalable streaming applications.

3.4 Data Integration

The most challenging task in Big Data research is to deal with the heterogeneity, diversity, and complexity of the data and to find better ways to integrate them. In addition to exploring the flexibility of NoSQL technology, another promising area is to apply ontologies and Semantic Web technology. As a formal, explicit specification of a shared conceptualization of a domain of interest, ontologies play an important role in addressing the issues of heterogeneity in data sources. Rapid development and adoption of ontologies have enabled the research community to annotate and integrate biological and biomedical data using standardized ontologies, and automate the discovery and composing of bioinformatics web services and workflows. Linked Data technology provides a method for publishing structured data on the web and making them interconnected. The successful Linked data projects in the field of bioinformatics include the Bio2RDF [187] and EBI RDF platforms [188]. They use Semantic Web technologies to build and provide the largest network of Linked data for the Life Sciences by defining a set of simple conventions to create RDF(s) compati-

ble Linked Data from a diverse set of heterogeneously formatted sources obtained from multiple data providers. The challenge for data integration using Linked Data is to develop applications that can consume such data, extract meaningful biological knowledge, and present it in a user-friendly fashion.

3.5 User Interfaces

With the pervasiveness of mobile devices (tablets and phones), responsive web design that makes the web page look good on all devices becomes more and more important. Next-generation protein bioinformatics databases should provide users with an optimal viewing and interaction experience across a wide range of devices using technology such as Bootstrap [189], JQuery [190], Dojo Toolkit [191], etc. The need for speed, particularly for web-based applications, has also driven the development of NoSQL technology and high-performance index and search platforms such as Lucene/Solr [192] for fast information retrieval.

4 Conclusions

In this chapter, we presented a comprehensive review (with categorization and description) of major protein bioinformatics databases. We also reviewed and discussed the recent technology improvements that can help addressing some of the challenges in building next-generation protein bioinformatics databases and resources in the Big Data era.

Acknowledgments

This work was supported by grants from the National Institutes of Health: U41HG007822 and P20GM103446.

References

1. Ridley M (2006) *Genome*. Harper Perennial, New York
2. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the yeast transcriptome. *Cell* 2:243–251
3. Anderson NL, Anderson NG (1998) Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* 11:1853–1861
4. Hye A, Lynham S, Thambisetty M, Causevic M, Campbell J, Byers HL, Hooper C, Rijdsdijk F, Tabrizi SJ, Banner S, Shaw CE, Foy C, Poppe M, Archer N, Hamilton G, Powell J, Brown RG, Sham P, Ward M, Lovestone S (2006) Proteome-based plasma biomarkers for Alzheimer's disease. *Brain* 11:3042–3050
5. Decramer S, Wittke S, Mischak H, Zürbig P, Walden M, Bouissou F, Bascands JL, Schanstra JP (2006) Predicting the clinical outcome of congenital unilateral ureteropelvic junction obstruction in newborn by urinary proteome analysis. *Nat Med* 4:398–400
6. Metzker M (2010) Sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46
7. Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, Chen Y, Wu CH

- (2011) A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* 27:1190–1191
8. Chen C, Huang H, Wu CH (2011) Protein bioinformatics databases and resources. *Methods Mol Biol* 694:3–24
 9. Farrell CM, O’Leary NA, Harte RA, Loveland JE, Wilming LG, Wallin C, Diekhans M, Barrell D, Searle SM, Aken B, Hiatt SM, Frankish A, Suner MM, Rajput B, Steward CA, Brown GR, Bennett R, Murphy M, Wu W, Kay MP, Hart J, Rajan J, Weber J, Snow C, Riddick LD, Hunt T, Webb D, Thomas M, Tamez P, Rangwala SH, McGarvey KM, Pujar S, Shkeda A, Mudge JM, Gonzalez JM, Gilbert JG, Trevanion SJ, Baertsch R, Harrow JL, Hubbard T, Ostell JM, Haussler D, Pruitt KD (2014) Current status and new features of the consensus coding sequence database. *Nucleic Acids Res* 42:D865–D872
 10. Kodama Y, Mashima J, Kosuge T, Katayama T, Fujisawa T, Kaminuma E, Ogasawara O, Okubo K, Takagi T, Nakamura Y (2015) The DDBJ Japanese genotype-phenotype archive for genetic and phenotypic human data. *Nucleic Acids Res* 43:D18–D22
 11. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Hoad G, Kanz C, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Plaister S, Sobhany S, Stoehr P, Vaughan R, Wu D, Zhu W, Apweiler R (2007) EMBL nucleotide sequence database in 2006. *Nucleic Acids Res* 35:D16–D20
 12. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister J, Bryant SH, Canese K, Clark K, DiCuccio M, Dondoshansky I, Federhen S, Feolo M, Funk K, Geer LY, Gorenkov V, Hoepfner M, Holmes B, Johnson M, Khotomlianski V, Kimchi A, Kimelman M, Kitts P, Klimke W, Krasnov S, Kuznetsov A, Landrum MJ, Landsman D, Lee JM, Lipman DJ, Lu Z, Madden TL, Madej T, Marchler-Bauer A, Karsch-Mizrachi I, Murphy T, Orris R, Ostell J, O’Sullivan C, Panchenko A, Phan L, Preuss D, Pruitt KD, Rubinstein W, Sayers EW, Schneider V, Schuler GD, Sherry ST, Sirotkin K, Siyan K, Slotta D, Soboleva A, Sousov V, Starchenko G, Tatusova TA, Trawick BW, Vakatos D, Wang Y, Ward M, Wilbur W, Yaschenko E, Zbicz K (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 43:D6–D17
 13. Pruitt KD, Tatusova T, Maglott DR (2006) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35:D61–D65
 14. The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43:D204–D212
 15. Pitarch A, Sánchez M, Nombela C, Gil C (2003) Analysis of the *Candida albicans* proteome. II. Protein information technology on the Net (update 2002). *J Chromatogr B Analyt Technol Biomed Life Sci* 787:129–148
 16. Zhou T, Zhou ZM, Guo XJ (2013) Bioinformatics for spermatogenesis: annotation of male reproduction based on proteomics. *Asian J Androl* 15:594–602
 17. Hoogland C, Mostaguir K, Sanchez JC, Hochstrasser DF, Appel RD (2004) SWISS-2DPAGE, ten years later. *Proteomics* 4:2352–2356
 18. Hoogland C, Mostaguir K, Appel RD, Lisacek F (2008) The World-2DPAGE constellation to promote and publish gel-based proteomics data through the ExPASy server. *J Proteomics* 71:245–248
 19. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res* 35:D786–D793
 20. Potenza E, Di Domenico T, Walsh I, Tosatto SC (2014) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res* 43:D315–D320
 21. Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, Khuri N, Spill YG, Weinkam P, Hammel M, Tainer JA, Nilges M, Sali A (2014) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 42:D336–D346
 22. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P, Hendrickx PM, Lagerstedt I, Mir S, Fernandez Montecelo MA, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Sanz-García E, Sen S, Slowley RA, Wainwright ME, Deshpande MS, Iudin A, Sahni G, Salavert TJ, Hirshberg M, Mak L, Nadzirin N, Armstrong DR, Clark AR, Smart OS, Korir PK, Kleywegt GJ (2015) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res* 44:D385–D395
 23. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H,

- Standley DM, Nakagawa A, Nakamura H (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res* 40:D453–D460
24. de Beer TA, Berka K, Thornton JM, Laskowski RA (2014) PDBsum additions. *Nucleic Acids Res* 42:D292–D296
25. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T (2013) The protein model portal—a comprehensive resource for protein structure and model information. Database. doi:10.1093/database/bat031
26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
27. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 31:3381–3385
28. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 35:D198–D201
29. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 42:D1083–D1090
30. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 42:D1091–D1097
31. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 42:D459–D471
32. Chang A, Schomburg I, Placzek S, Jeske L, Ulbrich M, Xiao M, Sensen CW, Schomburg D (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res* 43:D439–D446
33. Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Res* 28:304–305
34. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P (2014) The reactome pathway knowledgebase. *Nucleic Acids Res* 42:D472–D477
35. Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, Algaa E, Weidemann A, Sauer-Danzwith H, Mir S, Krebs O, Bittkowski M, Wetsch E, Rojas I, Müller W (2012) SABIO-RK—database for biochemical reaction kinetics. *Nucleic Acids Res* 40:D790–D796
36. Fazekas D, Koltai M, Türei D, Módos D, Pálffy M, Dúl Z, Zsákai L, Szalay-Bekő M, Lenti K, Farkas JJ, Vellai T, Csermely P, Korcsmáros T (2013) SignaLink 2—a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol* 7:7
37. Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, Bridge A, Bougueleret L, Xenarios I, Viari A (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res* 40:D761–D769
38. Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res* 34:D281–D284
39. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cucho BA, Bougueleret L, Poux S, Redaschi N, Xenarios I, Bridge A (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res* 43:D1064–D1070
40. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* 43:D213–D221
41. Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8:1551–1566
42. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) The Pfam protein families database. *Nucleic Acids Res* 42:D222–D230

43. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, Ledley RS, Suzek BE, Arminski L, Chen Y, Zhang J, Cardenas JL, Chung S, Castro-Alvarez J, Dinkov G, Barker WC (2004) PIRSE: family classification system at the Protein Information Resource. *Nucleic Acids Res* 32:D112–D114
44. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell A, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 31:400–402
45. Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, Peyruc D, Kahn D (2002) ProDom: Automated clustering of homologous domains. *Brief Bioinform* 3:246–251
46. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013) New and continuing developments at PROSITE. *Nucleic Acids Res* 41:D344–D347
47. Rappoport N, Karsenty S, Stern A, Linial N, Linial M (2011) ProtoNet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. *Nucleic Acids Res* 40:D313–D320
48. Letunic I, Doerks T, Bork P (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* 43:D257–D260
49. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J (2009) SUPERFAMILY—comparative genomics, datamining and sophisticated visualisation. *Nucleic Acids Res* 37:D380–D386
50. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O (2007) TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* 35:D260–D264
51. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M (2008) Bgee: integrating and comparing heterogeneous transcriptome data among species. *Lect Notes Comput Sci* 5109:124–131
52. Praz V, Jagannathan V, Bucher P (2004) CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Res* 32:D542–D547
53. Grennan AK (2006) Genevestigator. Facilitating web-based gene-expression analysis. *Plant Physiol* 141:1164–1166
54. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, McMurry J, Marioni JC, Malone J, Megy K, Rustici G, Tang AY, Taubert J, Williams E, Mannion O, Parkinson HE, Brazma A (2014) Expression atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* 42:D926–D932
55. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SM, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P (2015) Ensembl 2015. *Nucleic Acids Res* 43:D662–D669
56. Kersey PJ, Lawson D, Birney E, Derwent PS, Haimel M, Herrero J, Keenan S, Kerhornou A, Koscielny G, Kähäri A, Kinsella RJ, Kulesha E, Maheswari U, Megy K, Nuhn M, Proctor G, Staines D, Valentin F, Vilella AJ, Yates A (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res* 38:D563–D569
57. Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 33:D54–D58
58. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44:D457–D462
59. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJ, Yoo HS, Zhang C, Zhang Y, Sobral BW (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 42:D581–D591
60. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ (2015) The UCSC genome browser database: 2015 update. *Nucleic Acids Res* 43:D670–D681
61. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell

- KS, Christophides GK, Christley S, Dialynas E, Emmert D, Hammond M, Hill CA, Kennedy RC, Lobo NF, MacCallum MR, Madey G, Megy K, Redmond S, Russo S, Severson DW, Stinson EO, Topalis P, Zdobnov EM, Birney E, Gelbart WM, Kafatos FC, Louis C, Collins FH (2007) VectorBase: a home for invertebrate vectors of human pathogens. *Nucleic Acids Res* 35:D503–D505
62. Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Done J, Grove C, Howe K, Kishore R, Lee R, Li Y, Muller HM, Nakamura C, Ozersky P, Paulini M, Raciti D, Schindelman G, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Wong JD, Yook K, Schedl T, Hodgkin J, Berriman M, Kersey P, Spieth J, Stein L, Sternberg PW (2014) WormBase 2014: new views of curated biology. *Nucleic Acids Res* 42:D789–D793
63. Herzig V, Wood DL, Newell F, Chaumeil PA, Kaas Q, Binford GJ, Nicholson GM, Gorse D, King GF (2011) ArachnoServer 2.0, an updated online resource for spider toxin sequences and structures. *Nucleic Acids Res* 39:D653–D657
64. Inglis DO, Arnaud MB, Binkley J, Shah P, Skrzypek MS, Wymore F, Binkley G, Miyasato SR, Simison M, Sherlock G (2012) The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Res* 40:D667–D674
65. Kaas Q, Yu R, Jin AH, Dutertre S, Craik DJ (2012) ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res* 40:D325–D330
66. Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Wieggers TC, Mattingly CJ (2015) The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic Acids Res* 43:D914–D920
67. Basu S, Fey P, Pandit Y, Dodson RJ, Kibbe WA, Chisholm RL (2013) DictyBase 2013: integrating multiple Dictyostelid species. *Nucleic Acids Res* 41:D676–D683
68. Misra RV, Horler RS, Reindl W, Goryanin II, Thomas GH (2005) EchoBASE: an integrated post-genomic database for *Escherichia coli*. *Nucleic Acids Res* 33:D329–D333
69. Zhou J, Rudd KE (2013) EcoGene 3.0. *Nucleic Acids Res* 41:D613–D624
70. Combet C, Garnier N, Charavay C, Grando D, Crisan D, Lopez J, Dehne-Garcia A, Geourjon C, Bettler E, Hulo C, Mercier PL, Bartenschlager R, Diepolder H, Moradpour D, Pawlowsky JM, Rice CM, Trepo C, Penin F, Deléage G (2007) euHCVdb: the European hepatitis C virus database. *Nucleic Acids Res* 35:D363–D366
71. Aurrecochea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer ET, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Srinivasamoorthy G, Stoeckert CJ Jr, Thibodeau R, Treatman C, Wang H (2010) EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Res* 38:D415–D419
72. dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase Consortium (2015) FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* 43:D690–D697
73. Frézal J (1998) Genatlas database, genes and development defects. *C R Acad Sci III* 321:805–817
74. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V, Lancet D (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* 18:1542–1543
75. Lechat P, Hummel L, Rousseau S, Moszer I (2008) GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res* 36:D469–D474
76. Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, Amarasinghe V, Youens-Clark K, Thomason J, Preece J, Pasternak S, Olson A, Jiao Y, Lu Z, Bolser D, Kerhornou A, Staines D, Walts B, Wu G, D'Eustachio P, Haw R, Croft D, Kersey PJ, Stein L, Jaiswal P, Ware D (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res* 42:D1193–D1199
77. Yamasaki C, Murakami K, Takeda J, Sato Y, Noda A, Sakate R, Habara T, Nakaoka H, Todokoro F, Matsuya A, Imanishi T, Gojobori T (2009) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res* 38:D626–D632
78. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res* 41:D545–D552
79. Uhlén M, Björling E, Agaton C, Szgyarto CA, Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Asplund C, Berglund L, Bergström K, Brumer H, Cerjan D,

- Ekström M, Elobeid A, Eriksson C, Fagerberg L, Falk R, Fall J, Forsberg M, Björklund MG, Gumbel K, Halimi A, Hallin I, Hamsten C, Hansson M, Hedhammar M, Hercules G, Kampf C, Larsson K, Lindskog M, Lodewyckx W, Lund J, Lundeberg J, Magnusson K, Malm E, Nilsson P, Odling J, Oksvold P, Olsson I, Oster E, Ottosson J, Paavilainen L, Persson A, Rimini R, Rockberg J, Runeson M, Sivertsson A, Skölleremo A, Steen J, Stenvall M, Sterky F, Strömberg S, Sundberg M, Tegel H, Tourle S, Wahlund E, Waldén A, Wan J, Wernérus H, Westberg J, Wester K, Wrethagen U, Xu LL, Hober S, Pontén F (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* 4:1920–1932
80. Kikuno R, Nagase T, Nakayama M, Koga H, Okazaki N, Nakajima D, Ohara O (2004) HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE. *Nucleic Acids Res* 32:D502–D504
81. Moszer I, Glaser P, Danchin A (1995) SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology* 141:261–268
82. Kapopoulou A, Lew JM, Cole ST (2011) The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)* 91:8–13
83. Andorf CM, Cannon EK, Portwood JL, Gardiner JM, Harper LC, Schaeffer ML, Braun BL, Campbell DA, Vinnakota AG, Sribalasu VV, Huerta M, Cho KT, Wimalanathan K, Richter JD, Mauch ED, Rao BS, Birkett SM, Richter JD, Sen TZ, Lawrence CJ (2015) MaizeGDB 2015: New tools, data, and interface for the maize model organism database. *Nucleic Acids Res* 44:D1195–D1201
84. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, The Mouse Genome Database Group (2015) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res* 43:D726–D736
85. Biaudet V, Samson F, Bessières P (1997) Micado—a network-oriented database for microbial genomes. *Comput Appl Biosci* 13:431–438
86. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders. *Nucleic Acids Res* 33:D514–D517
87. Gaudet P, Argoud-Puy G, Cusin I, Duek P, Evalet O, Gateau A, Gleizes A, Pereira M, Zahn-Zabal M, Zwahlen C, Bairoch A, Lane L (2013) neXtProt: organizing protein knowledge in the context of human proteome projects. *J Proteome Res* 12:293–298
88. Aymé S, Schmidtke J (2007) Networking for rare diseases: a necessity for Europe. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 50:1477–1483
89. Thorn CF, Klein TE, Altman RB (2005) PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. *Methods Mol Biol* 311:179–191
90. Wood V, Harris MA, McDowall MD, Rutherford K, Vaughan BW, Staines DM, Aslett M, Lock A, Bähler J, Kersey PJ, Oliver SG (2012) PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res* 40:D695–D699
91. Winsor GL, Lo R, Ho Sui SJ, Ung KS, Huang S, Cheng D, Ching WK, Hancock RE, Brinkman FS (2005) *Pseudomonas aeruginosa* genome database and pseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res* 33:D338–D343
92. Shimoyama M, De Pons J, Hayman GT, Laulederkind SJ, Liu W, Nigam R, Petri V, Smith JR, Tutaj M, Wang SJ, Worthey E, Dwinell M, Jacob H (2015) The rat genome database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res* 28:D743–D750
93. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED (2012) *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Res* 40:D700–D705
94. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Plötz L, Singh S, Wensel A, Huala E (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40:D1202–D1210
95. Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList—10 years after. *Tuberculosis (Edinb)* 1:1–7
96. Bowes JB, Snyder KA, Segerdell E, Gibb R, Jarabek C, Noumen E, Pollet N, Vize PD (2008) Xenbase: a *Xenopus* biology and genomics resource. *Nucleic Acids Res* 36:D761–D767
97. Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C,

- Ramachandran S, Ruef BJ, Ruzicka L, Schaper K, Shao X, Singer A, Sprunger B, Van Slyke CE, Westerfield M (2013) ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Res* 41:D854–D860
98. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, Jensen LJ, von Mering C, Bork P (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* 42:D231–D239
99. Perrière G, Duret L, Gouy M (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res* 10:379–385
100. Duret L, Mouchiroud D, Gouy M (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res* 22:2360–2365
101. Sonnhammer EL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43:D234–D239
102. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32:D277–D280
103. Altenhoff AM, Škunca N, Glover N, Train CM, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, Gonnet GH, Dessimoz C (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res* 43:D240–D249
104. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* 41:D358–D365
105. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res* 42:D897–D902
106. Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Hériché JK, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R (2008) TreeFam: 2008 update. *Nucleic Acids Res* 36:D735–D740
107. Wu TJ, Shamsaddini A, Pan Y, Smith K, Crichton DJ, Simonyan V, Mazumder R (2014) A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database. doi:[10.1093/database/bau022](https://doi.org/10.1093/database/bau022)
108. Peterson TA, Adadey A, Santana-Cruz I, Sun Y, Winder A, Kann MG (2010) DMDM: Domain Mapping of Disease Mutations. *Bioinformatics* 26:2458–2459
109. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43:D470–D478
110. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res* 32:D449–D451
111. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42:D358–D363
112. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40:D857–D861
113. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43:D447–D452
114. Schaab C, Geiger T, Stoehr G, Cox J, Mann M (2012) Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol Cell Proteomics* 11:M111.014068
115. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C (2015) Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15:3163–3168
116. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R (2006) The PeptideAtlas project. *Nucleic Acids Res* 34:D655–D658
117. Vizcaino JA, Cote RG, Csordas A, Dianas JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim

- M, Contell J, O'Kelly G, Schoenegger A, Ovelleiro D, Perez-Riverol Y, Reisinger F, Rios D, Wang R, Hermjakob H (2013) The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 41:D1063–D1069
118. Wienkoop S, Staudinger C, Hoehenwarter W, Weckwerth W, Egelhofer V (2012) ProMEX—a mass spectral reference database for plant proteomics. *Front Plant Sci* 3:125
119. Duan G, Li X, Köhn M (2015) The human DEPhOsporylation database DEPOD: a 2015 update. *Nucleic Acids Res* 43:D531–D535
120. Ross KE, Arighi CN, Ren J, Huang H, Wu CH (2013) Construction of protein phosphorylation networks by data mining, text mining and ontology integration: analysis of the spindle checkpoint. Database doi:[10.1093/database/bat038](https://doi.org/10.1093/database/bat038)
121. Durek P, Schmidt R, Heazlewood JL, Jones A, Maclean D, Nagel A, Kersten B, Schulze WX (2010) PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res* 38:D828–D834
122. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res* 39:D261–DD27
123. Sadowski I, Breikreutz BJ, Stark C, Su TC, Dahabieh M, Raithatha S, Bernhard W, Oughtred R, Dolinski K, Barreto K, Tyers M (2013) The PhosphoGRID *Saccharomyces cerevisiae* protein phosphorylation site database: version 2.0 update. Database doi:[10.1093/database/bat026](https://doi.org/10.1093/database/bat026)
124. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E (2014) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 43:D512–D520
125. Campbell MP, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita KF, Lisacek F, Packer NH (2014) UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res* 42:D215–D221
126. The Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res* 43:D1049–D1056
127. Natale DA, Arighi CN, Blake JA, Bult CJ, Christie KR, Cowart J, D'Eustachio P, Diehl AD, Drabkin HJ, Helfer O, Huang H, Masci AM, Ren J, Roberts NV, Ross K, Ruttenberg A, Shamovsky V, Smith B, Yerramalla MS, Zhang J, AlJanahi A, Çelen I, Gan C, Lv M, Schuster-Lezell E, Wu CH (2014) Protein Ontology: a controlled structured network of protein entities. *Nucleic Acids Res* 42:D415–D421
128. Mari A, Rasi C, Palazzo P, Scala E (2009) Allergen databases: current status and perspectives. *Curr Allergy Asthma Rep* 9:376–383
129. Lombard V, Golaconda RH, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42:D490–D495
130. Lenfant N, Hotelier T, Velluet E, Bourne Y, Marchot P, Chatonnet A (2013) ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins: tools to explore diversity of functions. *Nucleic Acids Res* 41:D423–D429
131. Isberg V, Vroliing B, van der Kant R, Li K, Vriend G, Gloriam D (2014) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res* 42:D422–D425
132. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc MP (2006) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res* 34:D781–D784
133. Rawlings ND, Waller M, Barrett AJ, Bateman A (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* 42:D503–D509
134. Jeffery CJ (1999) Moonlighting proteins. *Trends Biochem Sci* 24:8–11
135. Murphy C, Powlowski J, Wu M, Butler G, Tsang A (2011) Curation of characterized glycoside hydrolases of fungal origin. Database. doi:[10.1093/database/bar020](https://doi.org/10.1093/database/bar020)
136. Fawal N, Li Q, Savelli B, Brette M, Passaia G, Fabre M, Mathé C, Dunand C (2013) PeroxiBase: a database for large-scale evolutionary analysis of peroxidases. *Nucleic Acids Res* 41:D441–D414
137. Roberts RJ, Vincze T, Posfai J, Macelis D (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* 43:D298–D299
138. Saier MH, Reddy VS, Tamang DG, Vastermark A (2014) The transporter classification database. *Nucleic Acids Res* 42:D251–D258
139. Frenkel-Morgenstern M, Gorohovski A, Lacroix V, Rogers M, Ibanez K, Boullosa C, Andres LE, Ben-Hur A, Valencia A (2013) ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res* 41:D142–D151
140. Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking of protein residues by importance. *J Mol Biol* 336:1265–1282
141. Good BM, Clarke EL, de Alfaro L, Su AI (2012) The Gene Wiki in 2011: community

- intelligence applied to human gene annotation. *Nucleic Acids Res* 40:D1255–D1261
142. Schmidt EE, Pelz O, Buhlmann S, Kerr G, Horn T, Boutros M (2013) GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Res* 41:D1021–D1026
143. Igarashi Y, Heureux E, Doctor KS, Talwar P, Gramatikova S, Gramatikoff K, Zhang Y, Blinov M, Ibragimova SS, Boyd S, Ratnikov B, Cieplak P, Godzik A, Smith JW, Osterman AL, Eroshkin AM (2009) PMAP: databases for analyzing proteolytic events and pathways. *Nucleic Acids Res* 37:D611–D618
144. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* 31:219–223
145. Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010. <https://www.ncbi.nlm.nih.gov/books/NBK25501/>
146. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R (2004) UniProt archive. *Bioinformatics* 20:3236–3237
147. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31:926–932
148. Chen C, Natale DA, Finn RD, Huang H, Zhang J, Wu CH, Mazumder R (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One* 6:e18910
149. Mostaguir K, Hoogland C, Binz PA, Appel RD (2003) The Make 2D-DB II package: conversion of federated two-dimensional gel electrophoresis databases into a relational format and interconnection of distributed databases. *Proteomics* 3:1441–1444
150. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10:980
151. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent WR, Yao H, Markley JL (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408
152. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics* 21:988–992
153. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R (2010) Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* 11:40–79
154. Dale JM, Popescu L, Karp PD (2010) Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics* 11:15
155. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240
156. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, Bardou P, Beck T, Blake A, Bonierbale M, Brookes AJ, Bucci G, Buetti I, Burge S, Cabau C, Carlson JW, Chelala C, Chrysostomou C, Cittaro D, Collin O, Cordova R, Cutts RJ, Dassi E, Di Genova A, Djari A, Esposito A, Estrella H, Eyraas E, Fernandez-Banet J, Forbes S, Free RC, Fujisawa T, Gadaleta E, Garcia-Manteiga JM, Goodstein D, Gray K, Guerra-Assunção JA, Haggarty B, Han DJ, Han BW, Harris T, Harshbarger J, Hastings RK, Hayes RD, Hoede C, Hu S, Hu ZL, Hutchins L, Kan Z, Kawaji H, Keliet A, Kerhornou A, Kim S, Kinsella R, Klopp C, Kong L, Lawson D, Lazarevic D, Lee JH, Letellier T, Li CY, Lio P, Liu CJ, Luo J, Maass A, Mariette J, Maurel T, Merella S, Mohamed AM, Moreews F, Nabihoudine I, Ndegwa N, Noirot C, Perez-Llomas C, Primig M, Quattrone A, Quesneville H, Rambaldi D, Reecy J, Riba M, Rosanoff S, Saddiq AA, Salas E, Sallou O, Shepherd R, Simon R, Sperling L, Spooner W, Staines DM, Steinbach D, Stone K, Stupka E, Teague JW, Dayem Ullah AZ, Wang J, Ware D, Wong-Erasmus M, Youens-Clark K, Zadissa A, Zhang SJ, Kasprzyk A (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 43:W589–W598
157. De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 34:W362–W365

158. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A (2015) ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 43:D1113–D1116
159. Haeussler M, Raney BJ, Hinrichs AS, Clawson H, Zweig AS, Karolchik D, Casper J, Speir ML, Haussler D, Kent WJ (2015) Navigating protected genomics data with UCSC Genome Browser in a Box. *Bioinformatics* 31:764–766
160. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. *Genome Res* 19:630–638
161. Adler BT, de Alfaro L, Kulshreshtha A, Pye I (2011) Reputation systems for open collaboration. *Commun ACM* 54:81–87
162. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* 9:345–350
163. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol* 25:894–898
164. Hermjakob H (2006) The HUPO proteomics standards initiative—overcoming the fragmentation of proteomics data. *Proteomics* 6:34–38
165. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 41:D456–D463
166. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 22:1459–1466
167. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5:976–989
168. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74:5383–5392
169. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75:4646–4658
170. Wein SP, Cote RG, Dumousseau M, Reisinger F, Hermjakob H, Vizcaino JA (2012) Improvements in the protein identifier cross-reference service. *Nucleic Acids Res* 40:W276–W280
171. Cote R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H (2010) The ontology lookup service: bigger and better. *Nucleic Acids Res* 38:W155–W160
172. Reisinger F, Martens L (2009) Database on demand—an online tool for the custom generation of FASTA formatted sequence databases. *Proteomics* 9:4421–4424
173. Hermjakob H, Apweiler R (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev Proteomics* 3:1–3
174. Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti R, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK Jr, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R (2004) A common open representation of mass spectrometry data and its application in a proteomics research environment. *Nat Biotechnol* 22:1459–1466
175. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S,

- Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human protein reference database-2009 update. *Nucleic Acids Res* 37:D767–D772
176. Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, Dana JM, De Las Rivas J, Dumousseau M, Galeota E, Gaulton A, Goll J, Hancock RE, Isserlin R, Jimenez RC, Kerssemakers J, Khadake J, Lynn DJ, Michaut M, O’Kelly G, Ono K, Orchard S, Prieto C, Razick S, Rigina O, Salwinski L, Simonovic M, Velankar S, Winter A, Wu G, Bader GD, Cesareni G, Donaldson IM, Eisenberg D, Kleywegt GJ, Overington J, Ricard-Blum S, Tyers M, Albrecht M, Hermjakob H (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods* 8:528–529
177. Torii M, Arighi CN, Li G, Wang Q, Wu CH, Vijay-Shanker K (2015) RLIMS-P 2.0: a generalizable rule-based information extraction system for literature mining of protein phosphorylation information. *IEEE/ACM Trans Comput Biol Bioinform* 12:17–29
178. Tudor CO, Ross KE, Li G, Vijay-Shanker K, Wu CH, Arighi CN (2015) Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. Database doi:[10.1093/database/bav020](https://doi.org/10.1093/database/bav020)
179. Cooper CA, Harrison MJ, Wilkins MR, Packer NH (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res* 29:332–335
180. von der Lieth CW, Freire AA, Blank D, Campbell MP, Ceroni A, Damerell DR, Dell A, Dwek RA, Ernst B, Fogh R, Frank M, Geyer H, Geyer R, Harrison MJ, Henrick K, Herget S, Hull WE, Ionides J, Joshi HJ, Kamerling JP, Leeﬂang BR, Lütteke T, Lundborg M, Maass K, Merry A, Ranzinger R, Rosen J, Royle L, Rudd PM, Schloissnig S, Stenutz R, Vranken WF, Widmalm G, Haslam SM (2011) EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology* 21:493–502
181. Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* 24:1214–1216
182. The OpenSFS and Lustre Community Portal. <http://lustre.opensfs.org>
183. The Apache Hadoop Project. <http://hadoop.apache.org>
184. The Apache Hive data warehouse software. <http://hive.apache.org>
185. The Apache Pig platform. <http://pig.apache.org>
186. The Apache Spark. <http://spark.apache.org>
187. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 41:706–716
188. Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A, Gehant S, Laibe C, Redaschi N, Wimalaratne SM, Martin M, Le Novère N, Parkinson H, Birney E, Jenkinson AM (2014) The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* 30:1338–1339
189. Bootstrap<http://www.getbootstrap.com>
190. JQuery<https://www.jquery.com>
191. Dojo Toolkit<https://dojotoolkit.org>
192. The Apache Lucene<http://lucene.apache.org>

Chapter 2

UniProt Protein Knowledgebase

Sangya Pundir, Maria J. Martin, and Claire O'Donovan

Abstract

The Universal Protein Resource (UniProt) is a freely available comprehensive resource for protein sequence and annotation data. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). Across the three institutes more than 100 people are involved through different tasks such as expert curation, software development, and support.

This chapter introduces the functionality and data provided by UniProt. It describes example use cases for which you might come to UniProt and the methods to help you achieve your goals.

Key words UniProt, Protein data, Protein tools

1 Introduction

The Universal Protein Resource (UniProt) is a freely available comprehensive resource for protein sequence and annotation data [1]. UniProt provides a number of datasets, the main ones being the UniProt Knowledgebase (UniProtKB), Proteomes, UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). An overview of these datasets can be seen in Fig. 1. UniProtKB is the central hub for all functional information on proteins [2]. It consists of two sections, the reviewed (Swiss-Prot) section contains expertly annotated entries and the unreviewed (TrEMBL) section contains computationally analyzed and annotated entries. UniRef provides clusters of UniProtKB sequences (including isoforms) based on sequence identity at resolutions of 100 % identity, 90 % identity, and 50 % identity. This helps compress sequence redundancy and speed up sequence similarity searches. UniParc is the sequence archive of all publicly available protein sequences, including those not part of the UniProtKB set. UniProt also provides the Proteomes dataset for species with completely sequenced genomes. A proteome is the set of proteins thought to be expressed by an organism. In addition to these core protein datasets, UniProt

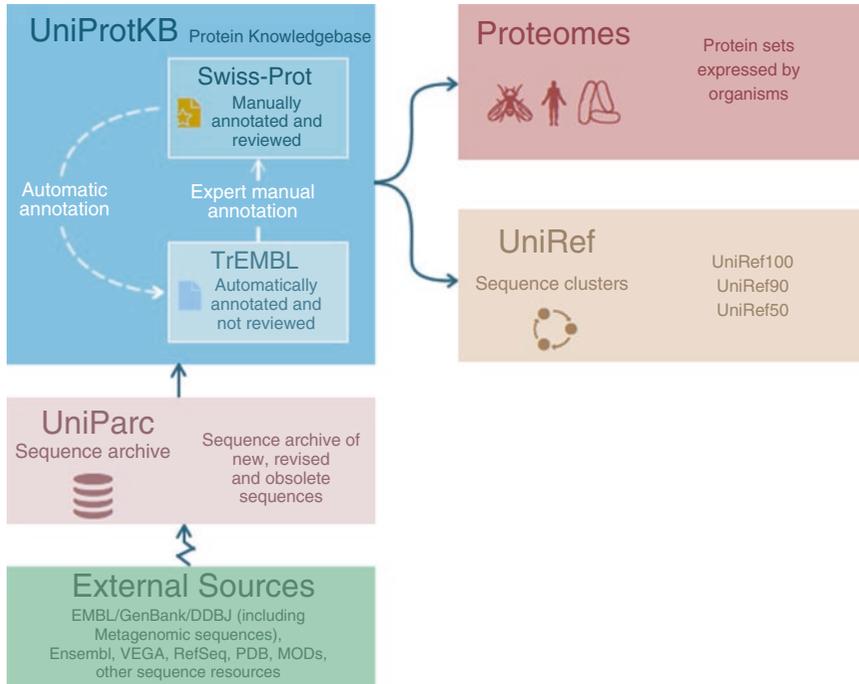


Fig. 1 Overview of key UniProt datasets

provides supporting datasets for Literature Citations, Taxonomy, Keywords, Subcellular locations, Cross-referenced databases, and Human diseases. UniProt also provides Automatic annotation rules for UniRule (expertly curated rules) and SAAS (statistical automatic annotation system).

UniProt also provides three tools embedded into workflow through datasets and on their own dedicated pages. These are the BLAST sequence search tool, the Align multiple sequence alignment tool, and the Retrieve/ID mapping tool that allows you to use a list of UniProtKB accessions to download a batch of UniProtKB entries or map the accessions to an external database and vice versa [3]. The tools are available through their own dedicated pages on the UniProt website at www.uniprot.org. They are also integrated into search results and entry pages from where you can access them while in the process of exploring data.

Understanding protein function is critical to research in many areas of science such as biology, medicine, and biotechnology. As the number of completely sequenced genomes continues to increase, huge efforts are being made in the research community to understand as much as possible about the proteins encoded by these genomes. This work is generating large amounts of data, which are spread across multiple locations including scientific literature and many biological databases. Keeping up with all of this information is a daunting task for most researchers and UniProt

supports this by providing a comprehensive body of protein information. Here, we describe the key use cases supported by UniProt for researchers to be able to achieve their goals at a single site.

2 Methods

2.1 Searching and Exploring Protein Data

The UniProt website provides an intuitive interface to help you find your protein of interest and explore protein data. You can use the search bar in the UniProt banner at the top of all pages to search the various UniProt datasets. Here, we consider searching for “insulin” as an example.

1. Go to <http://www.uniprot.org/>. You will see a drop-down list to the left of the search bar that allows you to select a dataset, *see* Fig. 2. You can search all UniProt datasets by selecting them from this drop-down. If you are looking for protein information about function, subcellular location, interactions, etc., use the default selection of “UniProtKB” and enter your search term in the search box (for example “insulin”) and click on the search button.
2. In order to make your search more specific, you can use the advanced search function. Click on the “advanced” link toward the right of the search box. Click on the dropdown in the advanced search panel to define the type of query you are making. For example, you can select “Protein name” from the first dropdown to correspond to the query “insulin.” You can also add additional parameters like “Organism,” “Protein existence,” etc., as shown in Fig. 3. To add more than two rows of parameters, click on the “+” icon and to delete a row of parameters, click on the bin icon. When you have entered all your parameters, click on the search button.
3. Once you have submitted your search, you will arrive at the relevant results page, for example the UniProtKB results page in Fig. 4. The results page offers a panel of filters on the left to help you refine your search. To the right of the filters is the main results table. Above the results table is a row of action

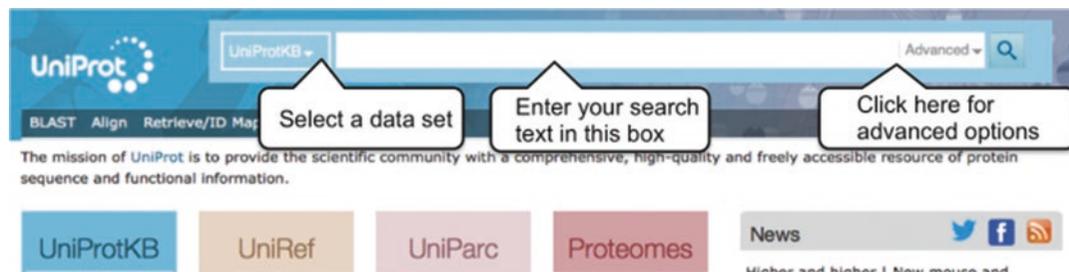


Fig. 2 UniProt header search bar

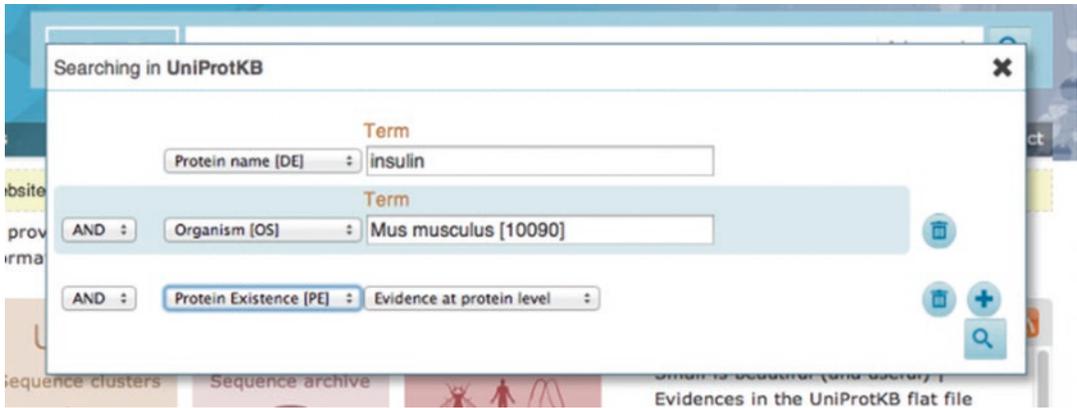


Fig. 3 Advanced search

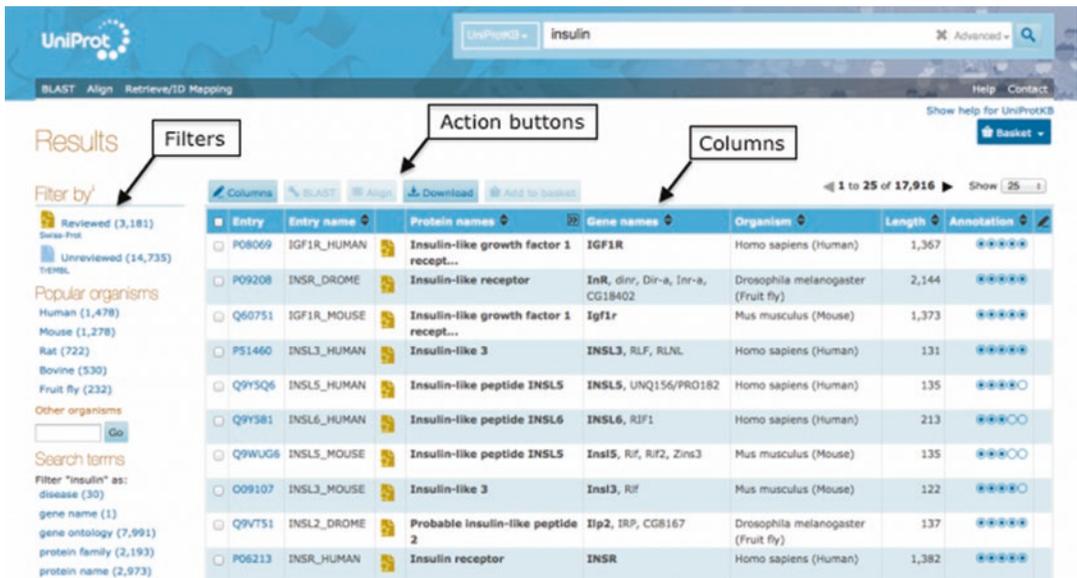


Fig. 4 UniProtKB search results page

buttons. You can select entries via checkboxes and then directly run a BLAST search, a multiple sequence alignment, download them in a number of available formats, or add them to your basket for later use (*see Note 1*). You can also edit the columns you are seeing to see more or less information by using the Columns button.

- When you have found your exact protein of interest in your dataset, you can click on the entry accession link highlighted in blue font to view the full protein entry page, as shown in Fig. 5. When viewing a UniProtKB entry, the menu bar on the left-hand side of the screen lists the entry sections, allowing you to move easily between sections. The entry provides all annotated

UniProtKB - P01308 (INS_HUMAN)

Display

Entry

Feature viewer

Feature table

Function

Names & Taxonomy

Subcellular location

Pathology & Biotech

PTM / Processing

Expression

Interaction

Structure

Family & Domains

Sequences (2)

Cross-references

Publications

Entry information

Miscellaneous

Similar proteins

BLAST Align Format Add to basket History

Feedback Help video Other tutorials and videos

Protein **Insulin**

Gene **INS**

Organism *Homo sapiens (Human)*

Status **Reviewed** - Annotation score: **5** - Experimental evidence at protein level¹

Function¹

Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver.

GO - Molecular function¹

- hormone activity @ Source: UniProtKB
- identical protein binding @ Source: JHAT
- insulin-like growth factor receptor binding @ Source: BHF-UCL
- insulin receptor binding @ Source: UniProtKB
- protease binding @ Source: UniProtKB

GO - Biological process¹

- activation of protein kinase B activity @ Source: BHF-UCL
- acute-phase response @ Source: BHF-UCL
- alpha-beta T cell activation @ Source: UniProtKB
- cell-cell signaling @ Source: UniProtKB
- cellular protein metabolic process @ Source: Reactome
- endocrine pancreas development @ Source: Reactome
- energy reserve metabolic process @ Source: Reactome
- ER to Golgi vesicle-mediated transport @ Source: Reactome

Fig. 5 UniProtKB protein entry

data for the protein, its sequence(s), and cross-references to over 150 relevant databases. You can also use action buttons on this page to run a BLAST search on the entry, align all isoforms (if any), view or download the entry in different formats, and add it to your basket for later.

UniProtKB entries also provide annotations about sequence features, such as domains, sites, PTMs, and variants. Colocalization of sequence features can have a significant impact on protein function (for example a variant at the position of an active site could alter enzyme function). To view all sequence features together, you can click on “Feature viewer” under the heading “Display” on the left-hand side of the page. It organises sequence features in category tracks, which can be expanded to view more detail, as shown in Fig. 6.

2.2 Finding the Proteome (Complete Protein Set) for an Organism

A proteome is the full set of proteins thought to be expressed by an organism and the UniProt websites provide proteomes datasets for species with completely sequenced genomes.

1. Click on the dropdown to the left of the search bar and select “Proteomes.”
2. Enter your query directly into the search box, for example “Homo Sapiens,” or click on the “advanced” button to the right of the search box and build a query using the parameters provided. This can help find exact results for the organism or taxonomy level you would like to specify. Click on the search button or hit enter to get to your results page.
3. You will be presented with a table of results for your proteomes search, as shown in Fig. 7.

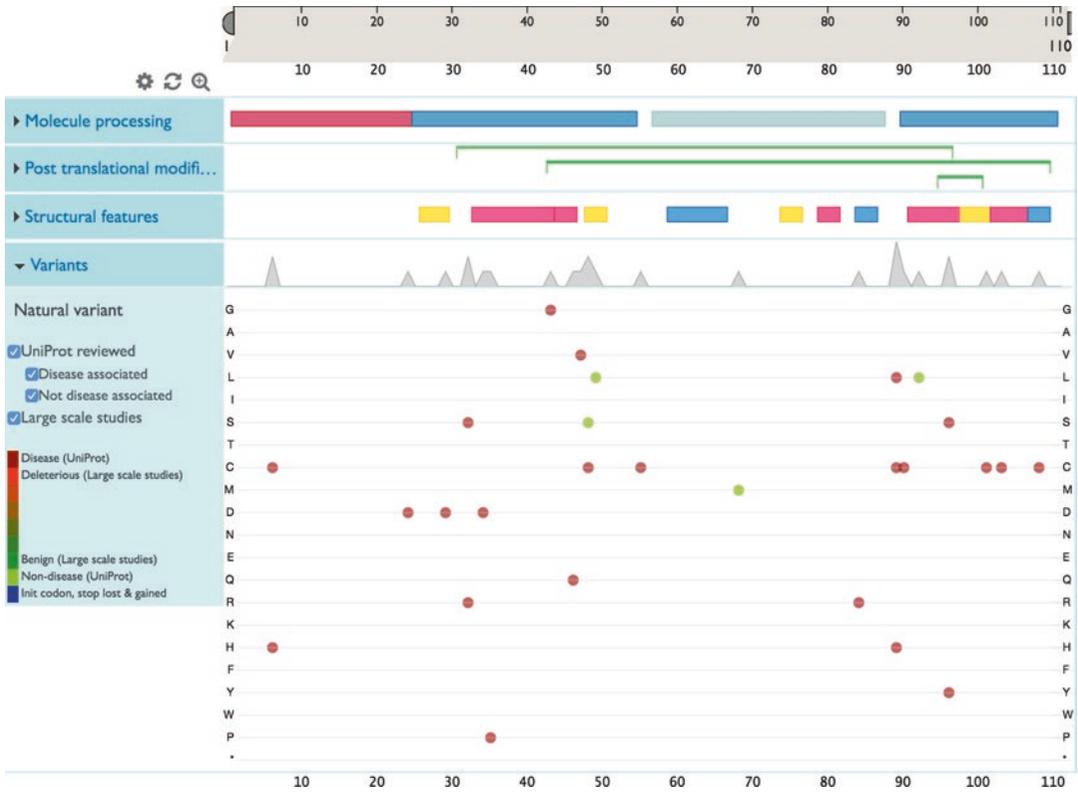


Fig. 6 UniProtKB protein feature viewer



Fig. 7 Proteomes results page

- Click on the proteome identifier to go to the detailed proteome page where you will see a summary of the organism, information about the genome assembly, proteins arranged by the chromosome or plasmid that they belong to, links to the protein entries in UniProtKB, and publications related to the proteome as shown in Fig. 8.

Proteomes - Homo sapiens (Human)

Display None

- Overview
- Components
- Publications

Overview

Proteome name	Homo sapiens - Reference proteome
Proteins	70,075
Proteome ID ¹	UP000005640
Taxonomy	9606 - Homo sapiens
Last modified	September 29, 2015
Genome assembly	GCA_000001405.19

Homo sapiens (*Homo sapiens sapiens*) or modern humans are the only living species of the evolutionary branch of great apes known as hominids. Divergence of early humans from chimpanzees and gorillas is estimated to have occurred between 4 and 8 million years ago. The genus *Homo* (*Homo habilis*) appeared in Africa around 2.3 million years ago and shows the first signs of stone tool usage. The exact lineage of *Homo* species is: *H. habilis*/*H. ergaster* to *H. erectus* to *H. rhodesiensis*/*H. heidelbergensis* to *H. sapiens* is still highly disputed. However, continuing evolution and in particular larger brain size and complexity culminates in *Homo sapiens*. The first anatomically modern humans appear in the fossil record around 200,000 years ago. Modern humans migrated across the globe essentially as hunter-gatherers until around 12,000 years ago when the practice of agriculture and animal domestication enabled large populations to grow leading to the development of civilizations. Overall life expectancy in Europe is 81 years.

Analysis of GRCh37 from Ensembl shows the human genome to contain 3.3 Gb and about 21,000 protein-coding genes and 196,000 gene transcripts.

Components¹

[Download](#) [View all proteins](#)

Component name	Genome accession(s)	Proteins
Chromosome 1	CM000663	5415
Chromosome 2	CM000664	4489
Chromosome 3	CM000665	4048
Chromosome 4	CM000666	2528
Chromosome 5	CM000667	2919
Chromosome 6	CM000668	1287

Fig. 8 Human Proteome entry

2.3 Finding All Proteins Involved in a Disease

Studying the involvement of proteins in diseases is important to help identify drug targets and better understand disease mechanisms in the human body. The best way to look for all proteins involved in a disease is to begin your search with the Human diseases dataset provided by UniProt [4]. Here, we consider this use case with the disease Breast Cancer as an example.

1. Go to <http://www.uniprot.org/> and click on the drop-down to the left of the search bar. Select “Human diseases” under the “Supporting data” section in the dropdown. Type your query into the search box, for example “Breast cancer,” and hit the search button.
2. You will arrive at a results page. The results table presents results that match your query such as is “Breast cancer, lobular,” “Breast cancer,” and so on. You will see a definition of the disease and a link to UniProtKB to see all proteins linked to this disease. For example, in case of the result “Breast cancer,” there are 12 linked UniProtKB entries as shown in Fig. 9.
3. To view all linked proteins, simply click on the UniProtKB link under the disease definition. You can also click on the disease name to view the detailed definition for the disease, its synonyms, and cross-references to related resources (like MIM, MeSH, etc.).
4. Clicking on the UniProtKB link will bring you to a UniProtKB results page for all proteins linked to this disease. Each of these UniProtKB protein entries provides information about various biological aspects such as function, taxonomy, subcellular

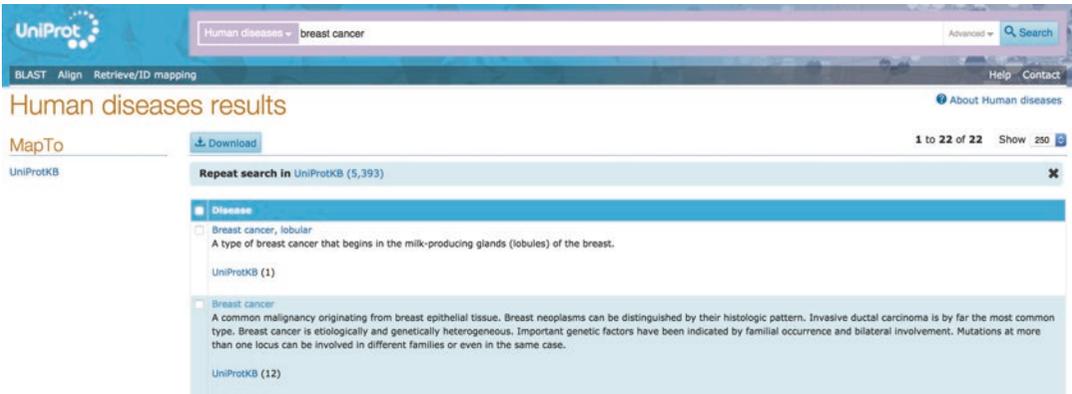


Fig. 9 Human diseases results page

location, pathology, biotech, etc. The “Pathology & Biotech” section of UniProtKB entries lists all diseases that the protein is involved in, the supporting evidence, and a list of natural variants linked to the disease.

2.4 Identifying Your Sequence Using the BLAST Search

If you have a protein sequence you would like to identify, you can use the Basic Local Alignment Search Tool to find closely matching sequences from UniProt that can help you understand evolutionary relationships and make functional inferences based on sequence identity. The UniProt website provides a form to submit your own sequences or any UniProtKB protein accession, UniParc sequence archive accession number, or UniRef cluster accession to the BLAST tool, using the NCBI BLAST algorithm [5]. It supports an integrated workflow that allows you to submit protein entries to BLAST from a search results page, the UniProt basket, and also a protein entry page.

1. Click on the BLAST link in the header of the UniProt website. This will bring you to the form submission page for BLAST.
2. Enter a protein or nucleotide sequence or a UniProtKB, UniParc, or UniRef cluster identifier or accession in the input field, for example P00750, as shown in Fig. 10.
3. You have a number of optional settings that you can change or leave as default. The options include “Target database,” “E-Threshold,” “Matrix,” “Filtering,” “Gapped” (yes or no) and number of Hits you’d like to get from the tool. For example, if you would like to find sequence matches only from a particular taxonomic level like “mammals” instead of from all of UniProtKB, you can select this from the “Target database” dropdown. You can also use the “Target database” dropdown to search against UniRef clusters instead of UniProtKB. UniRef clusters consist of UniProtKB sequences clustered based on identity at 100, 90, and 50 %. Searching against clusters hence speeds up BLAST searches.

BLAST

[About Blast](#)

P00750

Target databaseⁱ UniProtKB E-Thresholdⁱ 10 Matrixⁱ Auto Filteringⁱ None Gappedⁱ yes Hitsⁱ 250

Run Blast in a separate window.

[Clear](#) [Run BLAST](#)

Fig. 10 BLAST input page

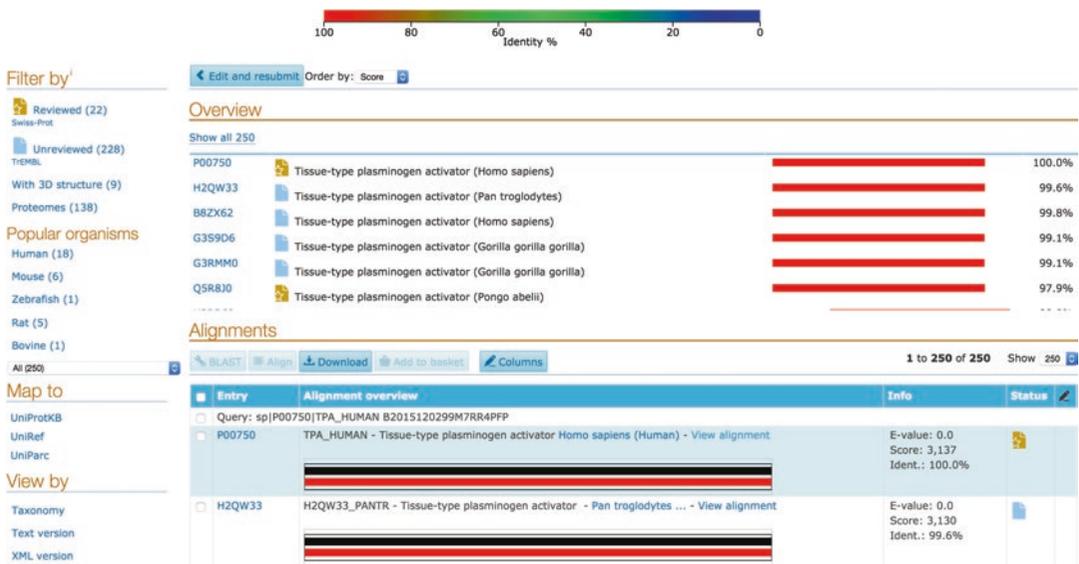


Fig. 11 BLAST results page

- Click on the Run BLAST button to execute your query. You will see a “Job status: RUNNING” page while your query is being run. This page provides details of your query sequence and settings.
- You will arrive at the BLAST results page once your query has been executed, as shown in Fig. 10. On the left-hand side, this page provides filters, mapping links to map your results to other datasets like UniProtKB, and alternative views of the results by taxonomy tree, text or XML versions. The upper half of the page provides an overview that you can expand to see all results by clicking on the “Show all 250” link. In Fig. 11, the overview shows the UniProtKB entry accession number, the protein names and species, a diagrammatic view of your matches that is

color coded by identity and the actual identity percentage. The lower half of the page shows your alignments in detail with each one represented diagrammatically in related to the query sequence. You can click on the graphic to view the raw sequence alignment in detail. The page also provides a Job identifier that you can use to retrieve your results page for up to 7 days.

6. You can also submit a UniProtKB, UniParc, or UniRef entry to the BLAST tool from a search results page by selecting the checkbox for that entry and clicking on the “BLAST” button above the results page. Alternatively, you can click on the checkbox and then click on the “Add to Basket” button above the search results table to build a collection of entries in your basket and submit one of them to BLAST at a later point.
7. When on a UniProtKB entry page, a UniRef cluster page, or a UniParc sequence archive page, you can simply click on the BLAST button near the top of the entry to submit the sequence to the BLAST tool. In case of a UniRef cluster entry with multiple sequences in the page, you can choose one by ticking on the checkbox to the left of it and then click on the “BLAST” button to submit to the tool.

2.5 Multiple Sequence Alignment

Aligning multiple sequences can help understand evolutionary relationships and identify areas of conservation between your sequences that can have structural or functional associations. UniProt provides a multiple sequence alignment tool “Align” that uses the Clustal Omega algorithm to align sequences [6]. For the most meaningful results, you should try and align sequences that are likely to be related so that you can explore evolutionary, structural, and functional relationships. You can access the tool through its own form submission page or directly through search results pages and protein entry pages. Integrating the tool into the data exploration workflow offers you a flexible way to find and analyze your data.

1. Click on the “Align” link in the header of the UniProt website. You will see a form submission page with an input box.
2. If you have two or more sequences that you would like to align to find areas to conservation and divergence, you can submit the sequences in FASTA format or accessions into the input box on this page. Click on the “Align” button to execute your query.
3. You will see a “Job status: RUNNING” page while your query is being executed.
4. Once completed, you will be presented the Alignment results page. The results page presents the alignment information, an evolutionary relationship tree for your sequences, and the results information at the bottom. On the left-hand side, you have Highlight options that allow you to select checkboxes to visually highlight sequence areas corresponding to

2.6 ID Mapping

If you have a list of UniProtKB accessions that you need to map or convert to identifiers from another database, for example if you have a list of UniProt accessions from a mass spectrometry experiment that you would like to map to other databases (for example, Ensembl, PDB, InterPro, etc.), you can use the Retrieve/ID mapping tool on the UniProt website. You can also map identifiers from external databases to UniProt using this tool [3].

1. Click on the “Retrieve/ID mapping” link the UniProt header. This will bring you to a form submission page with an input box, “from and to” database options and a “Go” button.
2. To convert UniProtKB accessions to an external database, for example Ensembl, paste your list of UniProtKB accessions in the input box or upload them as a file. Now click on the “From” dropdown and select UniProtKB and click on the “To” dropdown to select your target database (in this case Ensembl). The tool allows you to convert or map your accessions from UniProt to over 100 external databases that UniProt is cross-referenced to and vice versa (e.g., Ensembl, PDB, Refseq, etc.). You will get a results page showing a table of your input IDs and the mapped IDs from your target database, as shown in Fig. 13.
3. To convert external database identifiers to UniProtKB accessions or identifiers, for example Ensembl to UniProtKB, select the external database from the “From” dropdown and UniProtKB from the “To” dropdown. You will get a results page with your mapped UniProt entries and the default columns of data that you can customise, as shown in Fig. 14. You can select entries using checkboxes to BLAST them, align them, download them, or add them to your basket. You are also presented with filters on the left-hand side of the page to help narrow down your results.

Results



3 out of 4 identifiers from UniProtKB AC/ID were successfully mapped to 4 Ensembl IDs.

[Click here to download unmapped identifier\(s\)](#)



1 to 4 of 4

From	To
P31946	ENSG00000166913
P62258	ENSG00000108953
P62258	ENSG00000274474
ALBU_HUMAN	ENSG00000163631

1 to 4 of 4

Fig. 13 Retrieve/ID mapping results page from UniProt to external IDs

UniProtKB results

[About Upload lists](#)
[Basket](#)

4 out of 4 Ensembl identifiers were successfully mapped to 20 UniProtKB IDs in the table below.

Filter byⁱ

[BLAST](#)
[Align](#)
[Download](#)
[Add to basket](#)
[Columns](#)
1 to 20 of 20
Show 250

<input type="checkbox"/>	Your list: ...HMFIF	Entry	Entry name	Protein names	Gene names
<input type="checkbox"/>	ENSG00000163631	A0A087WWT3	A0A087WWT3_HUMAN	Serum albumin	ALB
<input type="checkbox"/>	ENSG00000163631	A0A0C4DGB6	A0A0C4DGB6_HUMAN	Serum albumin	ALB
<input type="checkbox"/>	ENSG00000163631	B7WNR0	B7WNR0_HUMAN	Serum albumin	ALB
<input type="checkbox"/>	ENSG00000163631	C9JKR2	C9JKR2_HUMAN	Albumin, isoform CRA_k	ALB hCG_149
<input type="checkbox"/>	ENSG00000163631	D6RCE7	D6RCE7_HUMAN	Serum albumin	ALB
<input type="checkbox"/>	ENSG00000163631	D6RHD5	D6RHD5_HUMAN	Serum albumin	ALB
<input type="checkbox"/>	ENSG00000163631	H0YA55	H0YA55_HUMAN	Serum albumin	ALB
<input type="checkbox"/>	ENSG00000163631	H7C013	H7C013_HUMAN	Serum albumin	ALB

Fig. 14 Retrieve/ID mapping results page from external database to UniProt ACs

- UniProt also provides the flexibility of submitting UniProt accessions to the ID mapping tool from your basket. Just add entries to your basket as you explore data and then you can use checkboxes to select them inside your basket and click on the “map Ids” tool to arrive on the Retrieve/ID mapping tool with your input prefilled in the input box.

2.7 Retrieving UniProt Entries for a List of Identifiers

If you have a list of UniProt accessions and would like to retrieve information for them from the UniProt website in a single step, you can use the Retrieve/ID mapping tool.

- Click on the “Retrieve/ID mapping” link the UniProt header. This will bring you to a form submission page with an input box, “from and to” database options and a “Go” button.
- To retrieve UniProtKB information corresponding to UniProtKB accessions or identifiers, paste your list of UniProtKB accessions in the input box or upload them as a file. You can leave the “From” dropdown and the “To” dropdown selections as the default UniProtKB since you are not converting or mapping identifiers between different databases.
- You will get a results page with your requested UniProt entries and the default columns of data that you can customise, as shown in Fig. 15. You can select entries using checkboxes to BLAST them, align them, download them, or add them to your basket. You are also presented with filters on the left-hand side of the page to help narrow down your results.

UniProtKB results

4 out of 4 UniProtKB AC/ID identifiers were successfully mapped to 7 UniProtKB IDs in the table below.

Filter by:

- Reviewed (7)
- Popular organisms
- Human (3)
- E. coli K12 (2)
- ECOS7 (1)
- ECOL6 (1)
- View by
- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

BLAST	Align	Download	Add to basket	Columns	1 to 7 of 7	Show 250
Your list: ZNBJV	Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/>	P31946	1433B_HUMAN	14-3-3 protein beta/alpha	YWHAH	Homo sapiens (Human)	246
<input type="checkbox"/>	P62258	1433E_HUMAN	14-3-3 protein epsilon	YWHAH	Homo sapiens (Human)	255
<input type="checkbox"/>	ALBU_HUMAN	ALBU_HUMAN	Serum albumin	ALB GIG20,GIG42,PRO0903,PRO1708,PRO2044	Homo sapiens (Human)	609
<input type="checkbox"/>	EFTU_ECOLI	POCE48	Elongation factor Tu 2	tufB b3980,JW3943	Escherichia coli (strain K12)	394
<input type="checkbox"/>	EFTU_ECOLI	POCE47	Elongation factor Tu 1	tufA b3339,JW3301	Escherichia coli (strain K12)	394
<input type="checkbox"/>	EFTU_ECOLI	POA6N3	Elongation factor Tu	tufA Z4697,ECs4190 tufB Z5553,ECs4903	Escherichia coli O157:H7	394
<input type="checkbox"/>	EFTU_ECOLI	POA6N2	Elongation factor Tu	tufA c4111 tufB c4935	Escherichia coli O6:H1 (strain CFT073 / ATCC 700928 / UPEC)	394

Fig. 15 Retrieve/ID mapping results page for batch UniProtKB entry retrieval

3 Note

1. UniProt provides a basket functionality to help you store your UniProt entries of interest and then analyze them, download them, or view them at a later point. The basket saves entries from UniProtKB, UniParc, and UniRef. You can add entries to the basket from search results pages of these three datasets or from their individual entry pages through the “Add to basket” button. The basket lets you select entries using checkboxes and submit them to the BLAST, Align, and ID mapping tools. You can also download your entries in formats like List, Text, FASTA, Tab-separated, Excel, GFF, and XML. It also provides a “Clear” button and a “Full view” button which shows you your saved entries in a full results screen where you can use filters and add or remove columns to the results table. The basket keeps your saved entries until you clear your browser cookies.

Acknowledgments

This work was supported by grant U41HG007822 from the National Institutes of Health.

References

1. The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43(Database issue):D204–D212. doi:10.1093/nar/gku989
2. Magrane M, The UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011. doi:10.1093/database/bar009

3. Huang H, McGarvey PB, Suzek BE et al (2011) A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* 27(8):1190–1191. doi:[10.1093/bioinformatics/btr101](https://doi.org/10.1093/bioinformatics/btr101)
4. Magrane M, Pundir S (2015) UniProt: exploring protein sequence and functional information. <http://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional>. Accessed 17 Dec 2015
5. Ladunga I (2009) Finding homologs in amino acid sequences using network BLAST searches. *Current protocols in bioinformatics/editorial board*, Andreas D Baxevanis[et al] Chapter 3:Unit 3.4. doi:[10.1002/0471250953.bi0304s25](https://doi.org/10.1002/0471250953.bi0304s25)
6. Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539. doi:[10.1038/msb.2011.75](https://doi.org/10.1038/msb.2011.75)

Tutorial on Protein Ontology Resources

**Cecilia N. Arighi, Harold Drabkin, Karen R. Christie, Karen E. Ross,
and Darren A. Natale**

Abstract

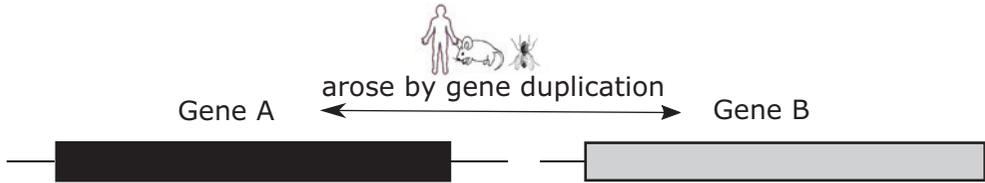
The Protein Ontology (PRO) is the reference ontology for proteins in the Open Biomedical Ontologies (OBO) foundry and consists of three sub-ontologies representing protein classes of homologous genes, proteoforms (e.g., splice isoforms, sequence variants, and post-translationally modified forms), and protein complexes. PRO defines classes of proteins and protein complexes, both species-specific and species non-specific, and indicates their relationships in a hierarchical framework, supporting accurate protein annotation at the appropriate level of granularity, analyses of protein conservation across species, and semantic reasoning. In the first section of this chapter, we describe the PRO framework including categories of PRO terms and the relationship of PRO to other ontologies and protein resources. Next, we provide a tutorial about the PRO website (proconsortium.org) where users can browse and search the PRO hierarchy, view reports on individual PRO terms, and visualize relationships among PRO terms in a hierarchical table view, a multiple sequence alignment view, and a Cytoscape network view. Finally, we describe several examples illustrating the unique and rich information available in PRO.

Key words Protein ontology, Proteoforms, Protein complexes, Post-translational modification, Orthologs

1 Introduction

Aberrations in protein activities are a fundamental cause of human diseases. Pathological changes in the proteome may result from single amino-acid variations resulting from nonsynonymous single nucleotide polymorphisms (nsSNPs) [1, 2], abnormal isoforms arising from aberrantly alternative spliced mRNAs [3, 4], changes in post-translational modifications (PTMs) [5, 6], or changes in cooperative behavior of multiple proteins in a protein complex [7, 8], as well as interdependencies of these mechanisms. With the advent of high-throughput proteomics technologies, our understanding of the protein composition of human cells in health and disease is expanding rapidly, especially when proteomics data are overlaid and analyzed along with genomic, transcriptomic, and interactomic data in their biological context.

a-FAMILY LEVEL:
all protein products from distinct genes related by common ancestry

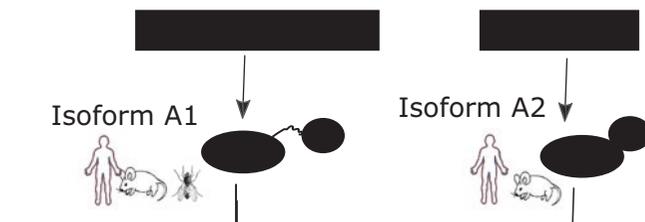


b-GENE LEVEL:
all protein products of a gene in a species and its 1:1 orthologs

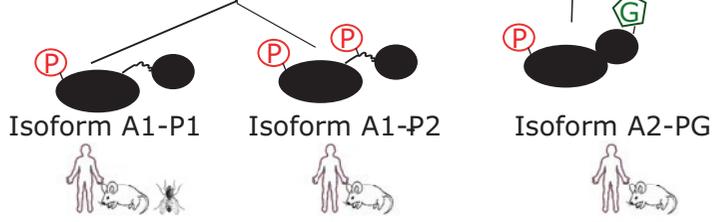


e.g., alternative splicing

c-SEQUENCE LEVEL:
distinct protein products upon initial translation



d-MODIFICATION LEVEL:
distinct proteoforms with post-translational modification/cleavage



e-COMPLEX LEVEL:
distinct protein complexes with specific components

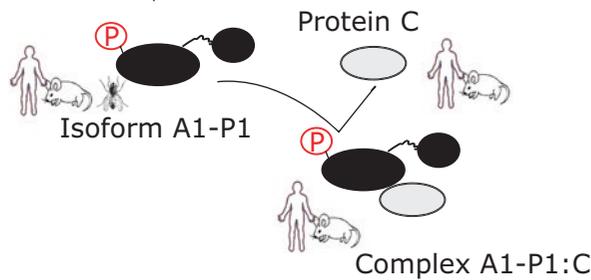


Fig. 1 PRO Hierarchy. The diagram shows the different classes that can be represented in the ontology. From *top to bottom*: a-Family class includes all the protein products of evolutionary-related genes at the homeomorphic level (e.g., all the protein products of Gene A and Gene B are in the same family class). These are conserved in a set of taxa (e.g., human, mouse, fly). b-Gene level includes all the protein products of a distinct gene in a species and its 1:1 orthologs. In PRO, human is the reference organism for vertebrates and all gene products of Gene A in human and its orthologs (e.g., mouse and fly) are in the same gene level class. Note that the gene is shown as a box because the gene structure (e.g., number and positions of introns) may differ between species. Species-specific organism-gene classes are children of

The Protein Ontology (PRO, proconsortium.org) [9] is a reference ontology for proteins and protein complexes in the OBO (Open Biological and Biomedical Ontologies) Foundry [10] that offers a research infrastructure for modeling biological systems and integrating existing and emerging experimental data.

PRO defines classes of proteins and protein complexes and indicates how these classes interrelate. For knowledge representation, PRO defines precise protein entities to support accurate annotation at the appropriate granularity and provides the ontological framework to connect all protein types necessary to model biology, in particular linking specific protein forms to particular complexes and particular functions in their biological context. For semantic data integration, PRO provides the ontological structure to connect—via specified relations—the vast amounts of protein knowledge contained in databases to support new hypothesis generation and testing.

Classes defined in PRO can be either organism non-specific or organism specific and range in granularity from protein family to proteoform classes (which account for the precise molecular form of a protein, including specification of sequence or splice variant and any post-translational modification or PTM [11]). Thus, it allows precise definition of protein objects and the specification of their relationships with each other.

1.1 PRO Framework

To model the various types of protein entities, we have formulated three sub-ontologies of PRO to represent: (1) protein classes of homologous genes, (2) protein forms (proteoforms [11]) arising from single genes, including splice isoforms, mutation variants, and PTM forms, and (3) protein complexes [9, 12, 13]. Protein terms in PRO are defined at multiple levels of granularity from the family level down to the isoform and/or modification level, allowing annotation at the most appropriate level given current knowledge. For example, as 14-3-3 proteins are encoded by several genes whose protein products may not be distinguishable in assays, they are represented by PR:000003237 for protein products of the 14-3-3 gene family. Similarly, when the protein is known to be the product of a given gene but the precise isoform is not known, then a gene-level PRO term covering all protein products is used (e.g., TP73, PR:O15350).

Figure 1 shows a schematic representation of the ontology, which is organized in different levels as follows:

Fig. 1 (continued) the corresponding Gene level (e.g., mouse Gene A, and human Gene A are both members of the same Gene A class). c-Sequence level includes all the isoforms produced by initial translation. This example shows two protein classes isoforms A1 and A2, created by alternative splicing, where isoform A1 is conserved in human, mouse, and fly species, and isoform A2 is observed only in mammals. Again, the species-specific organism-sequence terms can be created. d-Modification level includes all post-translational modifications. Shown here are proteoforms of isoform A1: P1 (phosphorylated at a single site) and P2 (phosphorylated at two sites) and proteoforms of Isoform A2: P1 (phosphorylated at a single site) and PG (phosphorylated and glycosylated). e-Protein complex level defines complexes based on component subunits (with stoichiometry if known). In this case, proteoform isoform A1 P1 and protein C are components of complex A1P1:C

- (a) *Family*: refers to the class of proteins translated from a specific set of ancestrally related genes. Proteins in this class can be traced back to a common ancestor showing homology over the entire length of the protein. The leaf-most nodes at this level are usually families comprising paralogous sets of gene products (of a single or multiple organisms). Figure 1 shows that gene A and gene B arose by gene duplication (paralogs) and that all protein products of gene A and gene B would be under the same family class in PRO. For example, in PRO the *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel protein class* (PR:000000676) is defined as “A protein with amino- and carboxyl-terminal intracellular domains separated by a domain (common with other ion channels) containing six transmembrane helices (S1–S6) in which the last two helices (S5 and S6) flank a loop, called the pore loop, which determines ion selectivity. The N terminus has a conserved domain that is also present in other voltage-gated potassium and sodium channels. The carboxyl-terminal region contains the cyclic nucleotide-binding domain (CNBD). In addition, there is a structural element called the C-linker, the region connecting the CNBD to the S6 segment, which couples conformational changes in the ligand-binding domain to channel activation ... [PMID:16382102]”. This class includes the protein products of genes HCN1, HCN2, HCN3, and HCN4.
- (b) *Gene*: a PRO term at this level refers to a class of proteins translated from a gene related by 1:1 orthology in distinct organisms. Considering human as a reference, all protein products of Gene A in human and its 1:1 orthologs in Fig. 1 would fall under the gene level class. The Gene A protein products from mouse and fly would also be included. Continuing with a real example, the HCN4 gene product (PR:000000708) is defined as “A potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel protein that is a translation product of the human HCN4 gene or a 1:1 ortholog thereof. [PRO:CNA].” This class currently includes the protein products of the HCN4 gene in rat, mouse, rabbit, and human. The species-specific genes (e.g., the human version of Gene A in Fig. 1, or mouse version of HCN4 in the example above) are children of the corresponding gene level terms. PRO uses the OMA orthology database [14] to map the organism gene to the corresponding gene level for selected model organisms.
- (c) *Sequence*: A PRO term at this level refers to the protein products with a distinct sequence upon initial translation. The sequence differences can arise from different alleles of a given gene, from splice variants of a given RNA, from alternative initiation or termination of transcription, and from ribosomal

frame shifting during translation. One can think of this as a mature mRNA-level distinction. In Fig. 1, isoform A1 (observed in human, mouse, and fly) and isoform A2 (observed only in human and mouse) would be two different classes in PRO. Again, under each term the species-specific terms can be created, and these are called ortho-isoforms.

- (d) *Modification*: a PRO term at this level refers to the protein products derived from a single mRNA species that differ because of some change (or lack thereof) that occurs after the initiation of translation (co- and post-translational). This includes sequence differences due to cleavage and/or chemical changes to one or more amino acid residues. The example in Fig. 1 shows two proteoforms of isoform A1 that differ in their phosphorylation state (single vs. doubly phosphorylated form). We have recently collected information about proteoforms of beta-catenin [15], specifically in relation to cancer. The phosphorylation state of beta-catenin influences its stability and its interacting partners. Phosphorylation of a specific set of residues is needed for its degradation. Phosphorylation of Ser-45 on human beta-catenin by casein kinase I (CKI) followed by the sequential phosphorylation of Thr-41, Ser-37, and Ser-33 phosphorylation by the glycogen synthase kinase, *GSK3B*, (PR:000035772) creates a recognition site for the ubiquitin ligase *BTRC*, which ubiquitinates beta-catenin, targeting it for degradation by the proteasome. Another modified proteoform of beta-catenin, phosphorylated on Tyr-654 (PR:000044478), has enhanced transcription-related functions. Fifteen proteoforms with distinct phosphorylation site combinations are described in PRO for human beta-catenin.
- (e) *Complex*: class of complexes with a specific defined subunit composition. PRO makes no distinction between complexes whose components are modified before or after complex formation. All complexes are grouped into the “complex” category regardless of the specific components.

1.2 PRO and Interoperability with Other Protein-Related Resources

PRO collaborates closely with other ontologies and resources to maximize semantic interoperability. The organism-specific protein complexes defined in PRO extend the generic complexes described in the Gene Ontology (GO) Cellular Component Ontology (GO, [16]). The organism-nonspecific complexes of GO provide parent terms for PRO’s organism-specific complex terms and provide the basis for connecting and comparing complexes between organisms. PSI-MOD [17] and the Sequence Ontology (SO) [18] are used to define protein classes of the modification category. PRO not only interoperates with ontologies, but with other resources as well. The “organism-gene” level in PRO is equivalent to the UniProtKB entries for the specific protein sequences, including

those of splice isoforms, arising from the gene represented [19]. Thus, PRO both incorporates UniProtKB and complements it by providing formal definitions for protein entities and placing the terms in an ontological context. Similarly, the Intact complex portal [20] contains protein complexes observed in specific major model organisms and these will be integrated into the ontological framework of PRO (ongoing effort).

1.3 PRO Applications

PRO has been employed in multiple studies including assisting in ontology building (especially in application ontologies), semantic integration, and functional annotation and ontological representation for proteoforms and complexes for proteomics studies. A few examples are listed below:

1. Used in combination with literature mining and curated databases to create a knowledge “map” for analysis of beta-catenin function in cancer [15].
2. Integrated in IDOBRU framework for ontological modeling of host-pathogen interactions using *Brucella* as the pathogen model [21].
3. Supporting GO annotation of specific proteoforms in model organism databases (e.g., in MGI [22] and PomBase [23]).
4. Supporting annotation of species-specific protein complexes in the Toll-like Receptor signaling pathway, relating both to their components and to species-independent families of complexes [12].
5. Used within the Neurological Disease Ontology, an ontology that represents aspects of neurological diseases that are relevant to their treatment and study [24].
6. Supporting concept recognition in CRAFT corpus [25].
7. Providing ontological framework for proteoforms in iPTM-net (<http://proteininformationresource.org/iPTMnet/>, see Chapter 16).

1.4 Scope of this Chapter

In this tutorial, you will learn how to use the PRO resources, including (1) searching/browsing the ontology and annotation; (2) analyzing proteoforms of a gene within and across species; (3) analyzing complexes; (4) saving/downloading data; and (5) visualization of the ontology. Examples in this book chapter are from Release 49.0.

2 Materials

1. The PRO website is accessible at proconsortium.org.
2. Download: The ontology (pro.obo) and the annotation (PAF.txt) can be downloaded from the ftp site accessible from the

```
[Term]
id: PR:000045512
name: kalirin isoform m8 phosphorylated 1 (mouse)
def: "A kalirin isoform m8 (mouse) that has been post-translationally modified to include phosphorylation at Ser-487. UniProtKB:A2CG49-8, Ser-487, MOD:00046." [PRO:KRC, PMID:22508986]
comment: Category=organism-modification. Evidence=(ECO:0000006, based on PMID:22508986).
Evidence=(ECO:0000181, based on PMID:22508986)
synonym: "Kal7 phosphorylated 1 (mouse)" EXACT []
is_a: PR:A2CG49-8 ! kalirin isoform m8 (mouse)
relationship: has_part MOD:00046 ! O-phospho-L-serine
relationship: only_in_taxon NCBITaxon:10090 ! Mus musculus
```

↙ ↘

Level Evidence code

Fig. 2 Example of a PRO OBO stanza

PRO home page. The ontology is also available in OBO and OWL formats through the OBO Foundry (7) and Bioportal (8). For general documentation, please see links on the PRO home page.

3. The pro.obo contains a stanza of information about each term. Each stanza in the obo file is preceded by [Term] and is composed of an ID, a name, synonyms (optional), a definition, comment (this section indicates the hierarchy level and may include evidence codes and other comments), cross-reference (optional), and one or more relationships to other terms (Fig. 2).
4. The annotations to PRO terms are distributed in the PAF.txt file. To facilitate interoperability to the best extent this tab delimited file follows the structure of the gene ontology association (GAF) file. Please read the README file and the PAF guidelines.pdf in the ftp site to learn about the structure of this file. PRO terms are annotated with relations to other ontologies or databases. Currently used: Gene ontology (GO) [16] to describe processes, function and localization; Sequence ontology (SO) [26] to describe protein features; PSI-MOD (<http://www.psiview.info/MOD>) to describe protein modifications; Disease Ontology[27] to describe disease states; and Pfam [28] to describe domain composition.
5. Link to PRO. PRO identifiers are URIs, globally unique identifiers. The URIs have one of two forms:
 - http://purl.obolibrary.org/obo/PR_ddddeeeeee where d is a digit.
 - http://purl.obolibrary.org/obo/PR_<uniprot accession>

For example, see http://purl.obolibrary.org/obo/PR_Q8CGC7

6. Cytoscape. PRO features a Cytoscape [29] view that provides an interactive and visual interpretation of PRO terms and their

relations. The view can be accessed via the Cytoscape icon displayed on PRO search results pages and entry reports. Cytoscape supports searches and customization of the layout and display; displays links to PRO entry pages, OBO stanzas, and annotations; and allows saving of the network image in PNG format.

3 Methods

3.1 PRO Homepage

The PRO homepage (<http://proconsortium.org/>) (Fig. 3) is the starting point of navigation through the protein ontology resources. The menu on the left side links to several documents and information pages, as well as to the ftp download page. The functionalities in the homepage include: (Subheading 3.1.1) PRO browser, (Subheading 3.1.2) PRO entry retrieval, (Subheading 3.1.3) text search, and (Subheading 3.1.4) annotation.

3.1.1 PRO Browser

The browser is used to explore the hierarchical structure of the ontology (Fig. 4). The icons with a plus and minus signs (Fig. 4a, 1) are for expanding and collapsing nodes, respectively. Next to these icons is a PRO ID, which links to the corresponding entry report, followed by the term name. Unless otherwise stated, the implicit relation between nodes is *is_a*. The number of terms to be displayed in the hierarchy page can be managed from the result page number box (Fig. 4a, 2). The column on the right (Fig. 4a, 3) shows the level within the hierarchy for each term. This column is customizable, and additional information can be added by selecting other information tabs. Finally, to retrieve all terms matching a particular keyword, enter the word or phrase in the Find box (Fig. 4a, 4) and all terms matching will be displayed in the hierarchy as shown in Fig. 4b for terms containing IRF3-P, the phosphorylated form of interferon regulatory factor 3.

The screenshot shows the PRO homepage interface. At the top left is the PRO Protein Ontology logo. At the top right is the National Institute of Genome Health and Biotechnology (NIGMS) logo with the NIH grant number SR01GM080646-02. Below the logos is a paragraph of text describing the PRO ontology. On the left side, there is a vertical menu with links: Consortium, Dissemination, PRO Wiki, Documentation, Downloads, PRO tutorial, PRO Publications, and PRO Statistics. The main content area contains three primary tools: 1. 'Browse PRO' with a 'Quick Browse' dropdown menu and an example 'methylated (sample output)'. 2. 'Retrieve a PRO entry (enter a PRO ID):' with a text input field and an example 'PR:000025934 (sample output)'. 3. 'Search PRO (enter text or ID):' with a text input field and an example 'smad (sample output)'. Below these tools is a link for 'Annotation: RACE-PRO PRO tracker'. Red circles with numbers 1 through 4 are overlaid on the interface to highlight these key features.

Fig. 3 PRO homepage. Main Functionalities include: (1) Browsing, (2) Retrieval, (3) Search, and (4) Annotation tools

A



PRO Hierarchy (Note that the implicit relationship is *is_a*, whereas *d* indicates *derives_from* relationship.)

27 shown of 203215 records | 1225 pages:

Navigation icons: back, forward, search, etc.

20+ per page

2

expand	sort (no)	sort (size)	find	Category
+	GO:0032991	macromolecular complex		
+	PR:000025493	LPS:GPI-anchored CD14 complex		complex
+	PR:000025494	LPS:secreted CD14 alpha		complex
+	GO:0043234	protein complex		
+	PR:000018263	amino acid chain		external
+	PR:000018264	proteolytic cleavage product		
+	PR:000000001	protein		
+	PR:000023236	(dimethylallyl)adenosine tRNA methyltransferase MiaB		gene
+	PR:P91416	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase homolog 1 (Caenorhabditis elegans)		organism-gene
+	PR:Q09407	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase homolog 2 (Caenorhabditis elegans)		organism-gene
+	PR:000003775	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase		gene
+	PR:000033822	1,2-epoxyphenylacetyl-CoA isomerase		gene
+	PR:000023475	1,2-phenylacetyl-CoA epoxidase, subunit A		gene
+	PR:000023476	1,2-phenylacetyl-CoA epoxidase, subunit B		gene
+	PR:000023477	1,2-phenylacetyl-CoA epoxidase, subunit C		gene
+	PR:000006106	1,25-dihydroxyvitamin D(3) 24-hydroxylase, mitochondrial		gene
+	PR:000023201	1,4-Dihydroxy-2-naphthoyl-CoA synthase		gene
+	PR:000035348	1,4-alpha-glucan branching enzyme GlgB		gene
+	PR:000007871	1,4-alpha-glucan-branching enzyme		gene
+	PR:000023200	1,4-dihydroxy-2-naaphthoate octaorenyltransferase		aene

B

14 shown of 203215 records

Navigation tabs: Synonym, Gene, MGI, HGNC, Pfam, PIRSF, Reactome, UniProtKB

expand	sort (no)	sort (size)	find	Category
	GO:0032991	macromolecular complex		
	GO:0043234	protein complex		
	PR:000025764	IRF3-P:IRF3-P complex	IRF3-P	complex
	PR:000025765	IRF3-P:IRF3-P complex (human)		organism-complex
	PR:000027098	IRF3-P:IRF3-P complex (mouse)		organism-complex
	PR:000027083	IRF3-P:IRF7-P complex		complex
	PR:000027084	IRF3-P:IRF7-P complex (human)		organism-complex
	PR:000036001	IRF3-P:IRF3-P:KPNA4 complex		complex
	PR:000036002	IRF3-P:IRF3-P:KPNA4 complex (human)		complex
	PR:000036042	IRF3-P:p65 complex		complex
	PR:000036043	IRF3-P:p65 complex (mouse)		organism-complex
	GO:0032993	protein-DNA complex		
	PR:000036046	IRF3-P:p65:ISRE complex		complex
	PR:000036047	IRF3-P:p65:ISRE complex (mouse)		organism-complex

Fig. 4 PRO Browser. Navigate the ontology through its hierarchical view. (a) In the PRO browser you can: (1) use plus and minus signs to expand/collapse terms, respectively; (2) change the number of terms displayed per page; (3) customize the information tabs; and (4) find terms matching a keyword or phrase. (b) PRO browser view with terms containing IRF3-P, using the “find” functionality

3.1.2 PRO Entry Pages

The PRO entry provides a report containing the ontological information and annotation available for a given PRO term. If you know the PRO ID you can use the “retrieve PRO entry” box in the homepage (Fig. 3 (2)) to get direct access to the report. Alternatively, you can open a report by clicking on the PRO ID listed in any page (e.g., from search or browser results).

The entry report may contain a subset or all of the following sections (Fig. 5):

1. *Ontology Information* (Fig. 5 (1)). This section is found in all entries. It displays the ontological information including term ID, name, synonyms, definition, comments, and parent term. A table with terms arranged by categories may be present in entries for terms that encompass many child classes (e.g., gene level,

This page represents a class of proteins encompassing all the protein products of the **EPRS** gene in **mouse**.

[Protein Forms](#) [Complex](#) [Annotations](#)

Ontology Information ①

PRO ID: PR:Q8CGC7 [Show OBO stanza / PAF](#)

PRO name: bifunctional glutamate/proline--tRNA ligase (mouse)

Synonyms: **PRO-short-label:** mEPRS
EXACT: bifunctional aminoacyl-tRNA synthetase (mouse)
RELATED: Qprs

Definition: A bifunctional glutamate/proline--tRNA ligase that is encoded in the genome of mouse. [[QMA:MOUSE01336](#), [PMID:15489334](#), PRO:HJD]

PRO Category: organism-gene

Parent: [PR:000007144](#) bifunctional glutamate/proline--tRNA ligase [Retrieve All terms OBO Stanza / PAF](#)

Organism-Specific	
Category	Number of Terms
organism-gene	1
organism-sequence	1
organism-modification	2
organism-complex	1

Term Hierarchy Visualization: [DAG](#) [Cytoscape](#)

Related Cross References ②

Db identifiers: [UniProtKB:Q8CGC7](#)

Interactive Sequence View ③ [Select/align proteoforms across species](#)

Modification - 🔍

• Number of sequence: 3 • Alignment length: 1512 • Scale: "-" = 19 amino acids

Protein Forms ④

PRO ID&Category	Complex	Annotation	Name	Short Label	Definition&Comment
PR:000007144 gene			bifunctional glutamate/proline--tRNA ligase	EPRS	A protein that is a translation product of the human EPRS gene or a 1:1 ortholog thereof.
PR:Q8CGC7 organism-gene			bifunctional glutamate/proline--tRNA ligase (mouse)	mEPRS	A bifunctional glutamate/proline--tRNA ligase that is encoded in the genome of mouse.
PR:000037785 organism-modification			bifunctional glutamate/proline--tRNA ligase phosphorylated 1 (mouse)	mEPRS/Phos:1	A bifunctional glutamate/proline--tRNA ligase (mouse) that has been phosphorylated on a Ser residue in the noncatalytic linker connecting the synthetase cores in an IFN-gamma-dependent manner. Example: UniProtKB:Q8CGC7-1 , Ser-999, MOD:00046 .
PR:000037786 organism-modification			bifunctional glutamate/proline--tRNA ligase unphosphorylated 1 (mouse)	mEPRS/UnPhos:1	A bifunctional glutamate/proline--tRNA ligase (mouse) that lacks phosphorylation on a residue analogous to Ser-999 in the amino acid sequence represented by UniProtKB:Q8CGC7-1 . Example: UniProtKB:Q8CGC7-1 , Ser-999, PR:000026291 .
PR:Q8CGC7-1 organism-sequence			bifunctional glutamate/proline--tRNA ligase isoform 1 (mouse)	mEPRS/iso:1	A bifunctional glutamate/proline--tRNA ligase isoform 1 that is encoded in the genome of mouse.

mEPRS forms found in complexes ⑤

mEPRS Component	Complexes
PR:000037785 mEPRS/Phos:1	PR:000037795 GAIT complex (mouse)

Functional Annotation ⑥

PRO Term	GO Annotation	PRO Centric View	GO Centric View
PR:000037785 mEPRS/Phos:1 Ser-999, MOD:00046	located_in GO:0097452 GAIT complex	MGI:5466658 , PMID:23071094	
	participates_in GO:0071346 cellular response to interferon-gamma	MGI:5466658 , PMID:23071094	
PR:000037786 mEPRS/UnPhos:1 Ser-999, PR:000026291	located_in GO:0017101 aminoacyl-tRNA synthetase multienzyme complex	MGI:5466658 , PMID:23071094	
	located_in NOT GO:0097452 GAIT complex	MGI:5466658 , PMID:23071094	

Fig. 5 PRO entry report for mouse Eprs. The report shows the following sections: (1) ontology information; (2) related cross references; (3) sequence viewer; (4) protein forms; (5) subunits in complexes; and (6) annotation

organism-gene level). This table provides a quick overview of the number of proteoforms and complexes related to a given entry, as is in the case for mouse Eprs gene in the example provided.

2. *Related Cross References* (Fig. 5 (2)). This section contains mappings to external databases that relate to the protein or complex report, such as UniProtKB in this example.

3. *Interactive Sequence View* (Fig. 5 (3)). The sequence viewer displays the protein sequence(s) defined in the entry with modified sites highlighted (color-coded based on each PTM). When the class includes more than one sequence (like our example that includes all products of mouse Eprs), a multiple sequence alignment is shown. Click on the magnifier glass to zoom in and explore specific sequence sections. The sequence viewer does not appear in protein complex reports.
4. *Protein Forms/Complex Subunits*. The Protein Forms section (found in protein reports) lists all the proteoforms related to the entry in a hierarchical way (Fig. 5 (4)). The Complex Subunits section (found in complex reports) lists all the proteoforms that are components of the protein complex. In our example, mouse Eprs has been observed as a phosphorylated form (on Ser-999) and in the corresponding unphosphorylated form. The numbers in the orange and green boxes next to the PRO ID indicate the presence of annotation or complex information for the particular term, respectively (*see steps 5 and 6* below).
5. *Forms found in complexes* (found in protein reports) (Fig. 5 (5)). This section lists all complexes that contain at least one proteoform described on the page. For example, the Ser-999 phosphorylated form (PR:000037785) of Eprs is a component of the mouse GAIT complex (PR:000037795). The green box next to PR:000037785 in the Protein Forms table indicates that it is found in a complex.
6. *Functional Annotation* (Fig. 5 (6)). Finally, this section shows the annotation of the term, including functional and disease information (source: PAF file). These annotations were contributed by the PRO consortium group and by community annotators through RACE-PRO (*see* Subheading 3.1.4). This table has two different views. The PRO-centric view displays the annotations for each PRO term. The annotations refer to different ontologies (e.g., GO and DO) as appropriate. On the other hand, the GO-centric view clusters all the terms that have a GO annotation in common. In that way you can see similarities among terms.

3.1.3 Searching PRO

Searching can be performed by entering a keyword or ID in the text Search PRO box on the right side of the homepage. For example, you can type the name of the protein for which you want to find related terms. Alternatively, the advanced search can be accessed by clicking on the [Search PRO](#) hyperlinked title above the text entry box (Fig. 3 (3)) on the home page. The advanced search page (Fig. 6) enables searches with Boolean operators (AND, OR, NOT), as well as null (not present)/not null (present) searches with several field options (*see Note 1*, Fig. 6 (1)).

The screenshot shows an advanced search interface. At the top, there is a "Quick Links" menu (2) and a "Batch Retrieval" link (3). Below this is a search form (1) with fields for "Taxon ID" (10090), "Category" (organism-modification), and "Ontology ID" (not null). There are also "add input box" and "del input box" buttons. Below the search form is a "Display Options" section (5) with "Fields Not in Display" and "Fields In Display" lists, and an "apply" button. At the bottom, there is a "Save Result As" option (6). The main part of the interface is a results table (4) with columns for PRO ID, PRO Name, PRO Term Definition, Category, Parent, Modifier, Relation, Ontology ID, protein Terms, Relative To, Annotation, Interaction With, Evidence Source, Evidence Code, Taxon ID, and Inferred From. Two results are visible, both for mouse proteoforms.

PRO ID	PRO Name	PRO Term Definition	Category	Parent	Modifier	Relation	Ontology ID	protein Terms	Relative To	Annotation	Interaction With	Evidence Source	Evidence Code	Taxon ID	Inferred From
PR-000025634	tyrosine-protein kinase SYK isoform 1 phosphorylated 1 (mouse)	A tyrosine-protein kinase SYK isoform 1 phosphorylated 1 in mouse. UniProtKB:P58025-1; Tyr-317/Tyr-242/Tyr-346; MOD:00048; [PMID:15536089; PRO:JAN]	organism-modification	PR-000025613	has_function	function	GO:0018901	protein kinase binding	PR-0002769	PR-0002769	PR-0002769	PMID:15536088	IP1	10090	
PR-000025892	transcription factor AP-1 isoform 1 phosphorylated 1 (mouse)	A transcription factor AP-1 isoform 1 phosphorylated 1 in mouse. UniProtKB:P55672-1; Ser-63/Ser-73; MOD:00046 [MOD:00696; PMID:12144319; PMID:20516211; PRO:HSD]	organism-modification	PR-000025891	has_function	function	GO:0003515	protein binding	PR-000238-1	PR-000238-1	PR-000238-1	PMID:20516211	IP1	10090	

Fig. 6 Advance search and result. (1) Search boxes with Boolean operators; (2) quick links to popular searches; (3) batch retrieval of terms using multiple identifiers; (4) result table; (5) display options to customize table content; and (6) saving option

Figure 6 (1) shows an example of advanced search intended to retrieve all PRO terms for mouse proteoforms (field->Taxon ID, “10090” and field->category, “organism-modification”) containing functional annotation (field->ontology ID, “not null”). In addition, the “Quick Links” menu (Fig. 6 (2)) gives direct access to popular searches (like searching for phosphorylated forms) and the “Batch Retrieval” link (Fig. 6 (3)) allows entering multiple identifiers (e.g., PRO and UniProt) in a single search.

In our search, 138 mouse proteoforms are shown in a results table (Fig. 6 (4)) with the following default columns: PRO ID, PRO name, PRO Term Definition, Category, Parent (term ID), and the searched fields. Some of the functionality in this page includes:

- Display Option (Fig. 6 (5)): Allows you to customize the result table by adding or removing columns. Use > to add or < to remove items from the list. Click the **apply** button for changes to take effect.
- Link to PRO entry reports: Clicking any hyperlinked PRO ID takes you to the corresponding PRO Entry report page.
- Link to hierarchical view: Clicking the blue hierarchy icon next to a PRO ID opens the browser with the selected term highlighted.
- Save Result As (Fig. 6 (6)): Allows you to save the result table as a tab-delimited file.

3.1.4 Annotation and PRO ID Requests

The annotation section is a forum for community interaction. There are two options: (1) the PRO tracker allows submission of new terms requests or changes/comments on existing ones, and (2) the rapid annotation interface, RACE-PRO, enables users to contribute directly to the curation of proteoforms.

The RACE-PRO interface can be used to:

1. Submit a request for a PRO ID for a proteoform of interest based on experimental evidence.
2. Add annotation to a proteoform or protein complex. Currently, in most databases, the annotation is added to the canonical protein. There is little or no distinction made between functions of isoforms or modified forms. Using RACE-PRO, the annotation can be associated with the most appropriate protein form. Therefore if a paper shows that only a phosphorylated form of isoform 2 of protein x is localized to the nucleus, then this annotation can be added only to PRO entry for the phosphorylated form of isoform 2, and not to others. Only experimental information is added. Another important consideration is that information submitted via RACE-PRO has to be pertinent to a particular protein sequence in a particular species.

How to Use RACE-PRO

In this section, we will demonstrate how to create the Ser-999 phosphorylated proteoform of mouse Eprs described in Subheading 3.1.2 and shown in the Protein Forms table in the entry report in Fig. 5, based on the information found in PMID:23071094. Before using RACE-PRO, the first step is to check if the proteoform is already in PRO by searching for the protein name or its UniProt accession in the search box on the home page. If it is not already in PRO, proceed to the RACE-PRO interface, as shown in Fig. 7. To access the RACE-PRO interface, it is necessary to fill in minimal personal information (name, e-mail address, institution) for the purpose of saving and accessing your data and for communication; this information will not be distributed to any third party or made publically available. The save option allows you to save your information to submit at a later time. Submit is used when you are done with the entry.

Definition of the protein object: In this block, enter all the information about a proteoform along with the evidence source.

1. *Retrieve the sequence:* Enter a UniProt accession to retrieve the relevant sequence. For example, for mouse Eprs, enter Q8CGC7 (*see Note 2*). The sequence will be displayed in the box on the RACE-PRO page (*see Note 3*).
2. *Specify sequence region:* allows selection of a subsequence in the case of cleaved products. After saving, the selected region will be underlined.



RACE-PRO
Rapid Annotation InterfaCE for PProtein Ontology (?)

[Save](#) [Submit](#)
Mon May 23 09:1

Annotator name:
 E-mail:
 Institution:

Note: Your e-mail address and other personal information are for internal use only and will not be shared with third parties.

Definition of the Protein Object

1. **UniProtKB identifier (?)**
 OR, click [here](#) to insert a different sequence:

```

-----+-----+-----+-----+-----+-----+
MAALCLTVNA GNPPEALLA VEHVKGDVSI SVEEGKENLL RVSETVAFTD VNSILRYLAR 60
IATTSGLYGT NLMHETEIDH WLEFSATKLS SCDRLTSAIN ELNHCLSLRT YLVGNLSLTA 120
DLCVWATLKG SAAWQEHLKQ NKTLLVHVKRW FGFLEAQQAF RSVGTKWDVS GNRATVAPDK 180
KQDVGKFVEL PGAEMGKVTV RFPPEASGYL HIGHAKAALL NQHYQVNFKG KLIMRFDOTN 240
PEKEKEDFEK VILEDVAMLH IKPDQFTYTS DHFETIMKYA EKLIQEGKAY VDDTPAEQMK 300
AEREQRTESK HRKNSVEKNL QMWEEMKKG SFGQSCCLRA KIDMSSNNGC MRDPTLYRCK 360
IQPHPRGTGNK YNVYPTYDFA CPIVDSIEGV THALRTTEYH DRDEQFYWII EALGIRKPYI 420
WEYSRLNLNN TVLSKRKLTV FVNEGLVDGW DDPFRPTVRG VLRGGMTEVG LKQFIAAQQS 480
SRSVVMEMWD KIWAFNKKVI DPVAPRYVAL LKKEVVPVNV LDAQEEMKEV ARHPKNPDVG 540
LKPVWYSPKV FIEGDAETF SEGEMVTFIN WGNINITKIH KNADGKITSL DAKLNLNENK 600
YKKTITITWL AESTHALSIP AVCVTEYHLI TKPVLGKDED FKQYINKDSK HEELMLGDPC 660
LKDLKKGDI IQLQRRGFFTC DQPYEPVSPY SCREAPCILI YIPDGHKTEM PTSGSKEKTK 720
VEISKKETS S APKERPAV SSTCATAEDS SVLYSRVAVG GDVVRELKAK KAPKEDIDAA 780
VKQLLTKAE YKEKTGQYK PGNPSAAAVQ TVSTKSSSNT VESTSLYNKV AAQGEVVRKL 840
KAEKAPKAKV TEAVECLLSL KAELYKKTGK DYVPGQPPAS QNSHSNPVSN AQPAGAEKPE 900
AKVLFDRVAC QGEVVRKLKA EKASKDQVDS AVQELLQLKA QYKSLTGIEY KPVSATGAED 960
KDKKKKEKEN KSEKQNKPKQ QNDGQKDS SSGSGLSSG GAGEGQPKK QTRLGLEAKK 1020
EENLAEWYSQ VITKSEMIY YDVSGCYILR PWSYSIWESI KDDFFDAEIKK LGVENCYFPI 1080
FVSQAALKEE KNHIEDFAPE VAWVTRSGKT ELAEPPIAIR TSETVMYPAY AKWVQSHRDL 1140
PVRLNQWCVN VRWEFKHPQ FLRTREFLWQ EGHSAFATFE EAADEVLLQIL ELYARVYEEEL 1200
LAIPVVRGRK TEKEKFPAGD YTTTIEAFIS ASGRAIQGAT SHHLLQNFSS MCEIVFEDPK 1260
TPGKQFAYQ CSWGLTTRTI GVMVMVHGDN MGLVLPPrVA SVQVVVPCG ITNALSEEDR 1320
EALMACNEY RRRLLGANIR VRVDLRDNYS PGWKFHWEL KGVVPRLEVQ PRDMKSCQFV 1380
AVRRDTGEKL TIAEKEAEAK LEKVLEDIQL NLFTRASEDL KTHMVVSNL EDFQKVLDAQ 1440
KVAQIPFCGE IDCEDWIKKM TARQDVEPG APSMGAKSLC IPFNPLCELQ PGAMCVCGKN 1500
PAKFYTLFGR SY

```

Organism:

2. Specify sequence region
 Full-length Region: from to

3. Indicate post-translational modifications (add amino acid number relative to the sequence displayed in the box 1) [\[more\]](#)
 Amino acid number: Phosphorylation Modifying enzyme:

4. Protein object name (separate multiple names using ";")

5. Evidence Source (separate multiple IDs using ";") [\[more\]](#)
 Db name: IDs: Evidence code:

6. Assay Evidence
 In vitro In vivo

Annotation of the Protein Object

Domain [\[add\]](#) [Link to PFAM](#)

Functional Annotation [\[add\]](#) [Link to GO](#)

Modifier	Relation	GO ID	GO term	Interaction with	Relative to	PMIDs
x	located_in	GO:00974:	GAIT complex			23071094

Fig. 7 RACE-PRO annotation interface

3. *Indicate post-translational modifications*: to describe a modification, or multiple cooccurring modifications, enter the residue number and the type of modification (*see Note 4*). The residue number should always refer to the sequence displayed in the sequence box. After saving, the residue(s) will be highlighted (in this example residue: 999, modification: phosphorylation). Check that the highlighted residues are in the expected positions. If there is no information about any post-translational modification, then leave these field blanks. It is also possible to indicate the modifying enzyme (e.g., kinase) if such information is available.
4. *Protein object name*: add names by which this object is referred to in the paper or source of data. By default the protein name in the UniProt record is displayed. Additional synonyms can be added separated by semicolons (;).
5. *Evidence Source*: Enter information about the source of the proteoform information. In this example, select PMID from the drop-down menu and add the ID 23071094. If the appropriate option is not present in the drop-down menu, use the “Other” option. In addition, the Evidence code menu is used to select if the information is experimental, based on similarity to another proteoform or deduced by the user based on a combination of sources and knowledge.
6. *Assay evidence*: use this to indicate if the data is from in vivo or in vitro experiments.

Annotation of the protein object: In this block annotation from experimental data that is pertinent to the protein form described in the previous section should be added. All the information about the different columns in the table is described in the PAF guidelines. This section is optional.

What Happens Next?

An editor from the PRO team will review the entry and request any additional information if needed. The corresponding PRO term will be generated along with associated annotations (if submitted). These will have the corresponding source attribution.

3.2 Interesting Examples

3.2.1 Proteoforms with Common Annotation

14-3-3 proteins are a family of proteins that bind to phosphorylated proteins and can affect the function of the target protein in many ways including the modulation of its enzyme activity, its sub-cellular localization, its structure and stability, or its molecular interactions [30]. To identify proteins that are regulated by these important modulators, search for terms that are annotated with the GO term “14-3-3 protein binding” (GO:0071889). Select the search field “Annotation term” and enter “14-3-3 protein binding” in the box. Some examples are listed in Table 1. As expected,

Table 1
Proteomeforms with 14-3-3 protein binding annotation (partial results)

PRO ID	PRO short	Relation	Ontology term (ID)	Interaction with	Evidence source
PR:000025725	HSF1-pSer303/pSer307	has_function	14-3-3 protein binding (GO:0071889)	PR:P62258	PMID:12917326
PR:000025837	MDM4-pSer367/ub			PR:P27348; PR:P61981	PMID:16511560; PMID:16511572
PR:000026140	BCL2-pSer75/pSer-99			PR:P63101	PMID:16932738
PR:000026848	RFWD2-pSer387			PR:O70456	PMID:20843328
PR:000029002	HDAC4-pS210/pS246				PMID:17179159
PR:000029006	HDAC5-pSer259/pSer498			PR:000044507; PR:P63104	PMID:11114197
PR:000044506	ABL1-pThr735			PR:P63104	PMID:15696159; PMID:16888623
PR:000044510	CTNNB1-pSer552			PR:P63104	PMID:17287208
PR:000044814	CSF2RB-pSer601			PR:P63104	PMID:10477722
PR:000044815	FBXO4-pSer12			PR:P62260	PMID:21242966
PR:000044817	PACS2-pSer437			PR:000044507	PMID:19481529
PR:000044818	PLK1-pSer99			PR:P61981	PMID:23695676
PR:000044819	PAK 4-pSer99/pSer474			PR:P61981	PMID:23695676
PR:000027894	FOXO1-pSer249	NOT_has_function	14-3-3 protein binding (GO:0071889)		PMID:18356527

all the proteins listed are phosphorylated forms, although they may contain other modifications as well. The specific 14-3-3 binding partner is listed in the “interacts with” column. However, note that the last entry, a phosphoproteoform of FOXO1, is explicitly annotated as NOT binding to 14-3-3 proteins. Therefore, this last example should be excluded in the final list of the phosphoproteoforms binding to 14-3-3 proteins.

3.2.2 Proteoform Conservation Across Species

The appetite-regulating hormone or Ghrelin is an endogenous peptidic hormone regulating both hunger and adiposity [31]. Acylated ghrelin induces a positive energy balance, while deacylated ghrelin has been reported to be devoid of any endocrine activities [32]. The hierarchy for this protein in PRO can be found at http://www.proconsortium.org/cgi-bin/pro/browser_pro?ids=PR:000007973#O and it reveals two isoforms: PR:000043841 and PR:000043842, with ortho-isoforms from human, mouse, and rat as child terms. Moreover, the mouse and human active cleaved acylated forms of Ghrelin are also conserved (PR:000044483 and PR:000044484, respectively) and both are children of the organism nonspecific modification term (PR:000044482).

Following the previous example of the mouse Eprs, the Interactive Sequence View (Fig. 5 (3)) can be used to check the proteoform conservation between mouse and other species. The link on the right upper corner of the viewer expands the sequence viewer to include other related sequences in PRO. Figure 8 shows the multiple alignment, in which the combination of modified residues in each proteoform is highlighted. This particular example points to differences between human and other species: human EPRS has two serine residues that can be phosphorylated (shown in gray in PR:P07814) whereas in other species the first one is not conserved (e.g., it is asparagine in mouse).

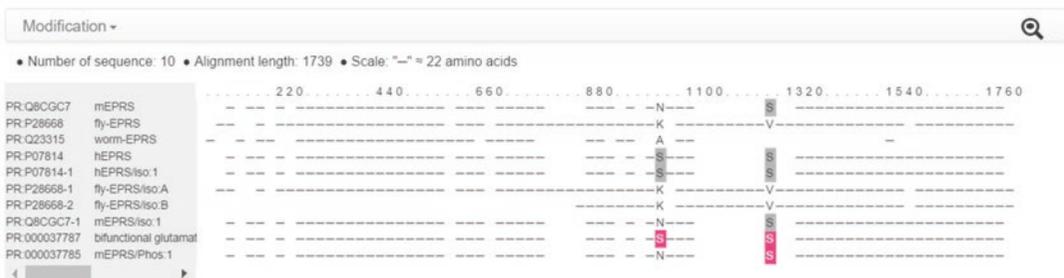


Fig. 8 Sequence viewer for mouse Eprs proteoforms and related PRO terms. Multiple sequence alignment with highlighting of the combination of modified residues in each proteoform (*pink*: experimentally shown to be phosphorylated; *gray*: phosphorylation site conserved)

3.2.3 Find Proteoforms and Complexes Associated with Disease

PRO includes annotations of proteoforms or complexes related to disease. These annotations are connected to disease ontology terms via two relations: `associated_with_disease_progression` and `associated_with_disease_suppression`. You can retrieve all the proteins and complexes that have been annotated in the ontology with “`associated_with_disease_progression`,” by searching for this relation in the ontology (select the search field “Relation” and enter “`association_with_disease_progression`” in the box), or you could find associations for a particular disease (select the search field “Ontology term” and enter a disease name, e.g., “cancer”).

3.2.4 Comparing Protein Complexes

The mitochondrial isocitrate dehydrogenase complex (NAD⁺) catalyzes the oxidative decarboxylation of isocitrate and it is important for the regulatory control of mitochondrial energy metabolism. If you search PRO using “mitochondrial isocitrate dehydrogenase complex (NAD⁺)” you will retrieve all the complexes currently in PRO, including a human, a yeast, and three mouse complexes. All of these species-specific complexes are children of the complex term GO:0005962 mitochondrial isocitrate dehydrogenase complex (NAD⁺). To review similarities and differences between the different complexes, select all the checkboxes and then click on Cytoscape link. The graphical view will open in a separate window. Figure 9 shows the complexes (squares) and their protein components (circles) connected by the relation `has_component` represented by the dashed arrows (*see Note 5*). The parent GO complex term is defined in terms of the enzymatic activity (Mitochondrial complex that possesses isocitrate dehydrogenase (NAD⁺) activity). The Cytoscape view of the organism-specific complexes reveals the different composition in yeast versus the human and mouse. While the yeast complex is composed of

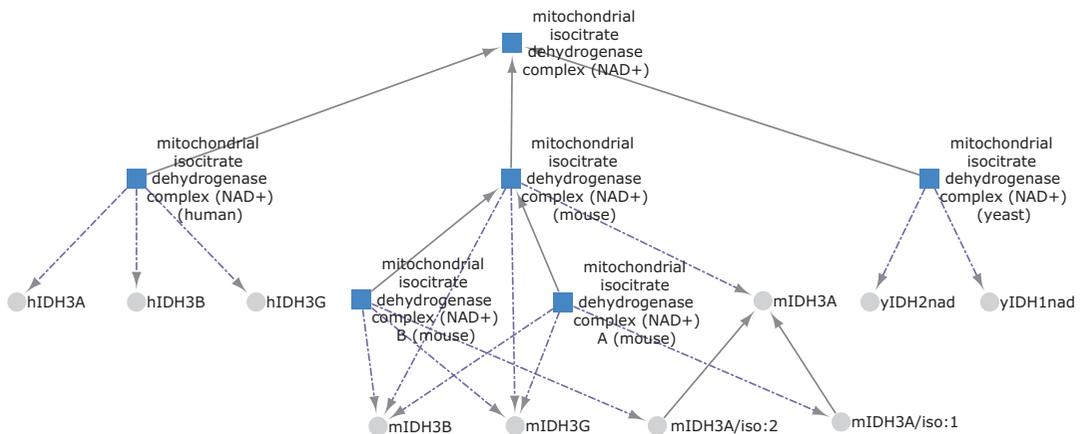


Fig. 9 Cytoscape view for the mitochondrial isocitrate dehydrogenase complex (NAD⁺) in PRO. Complexes are represented by *squares*, proteins by *circles*. *Solid arrows* represent the `is_a` relation, while *broken arrows* represent the `has_component` relation

two distinct proteins, the human and mouse complexes all have three. Selecting a node provides the definition of the term and also links to its report. The relationship between the three complexes for mouse is clearly show, with one being the parent of the other two that differ with respect to which isoform of the alpha subunit is present.

Another rich example is provided by complexes of cyclin-dependent kinases (CDKs) with cyclins. CDKs are a family of multifunctional enzymes that can modify various protein substrates involved in cell cycle progression. Their activity is controlled by their phosphorylation state and their binding to a cyclin regulatory subunit. All CDK:cyclin complexes in PRO are under cyclin-dependent protein kinase holoenzyme complex (GO:0000307). The hierarchical view of this term in PRO (http://www.proconsortium.org/cgi-bin/pro/browser_pro?ids=GO:0000307) lists all the complexes of CDKs and cyclins that have been curated. The Cytoscape view can provide a more granular level where the specific subunits, including the modifications, can be displayed. For example, PRO contains five different “cyclin B1:cdk1 complex” for humans which differ in the phosphorylation state of its component. Thus, the Cytoscape view in PRO facilitates the comparison of complexes within and across species.

4 Notes

1. Search tips:
 - (a) To retrieve all the entries from a given category, for example, all the nodes for gene product level, search by selecting the field “category” and entering “gene” in the box.
 - (b) Some of the search fields are of the type null/not null. This is the case for the ortho-isoform and ortho-modified form. To retrieve the ortho-isoform entries, select the search field ortho-isoform and type not null.
 - (c) More details about the options for the DB ID, Modifiers, and Relations fields are listed in the PAF guidelines (*see* Subheading 2).
2. To search UniProt accessions at the UniProt website (www.uniprot.org), enter the protein name and organism: eprs and mouse into the search box. From the result list, check the one that is relevant to your search. Alternatively, enter eprs and then use the filter to select mouse as the organism. View likely UniProt entries to confirm that it represents the protein of interest. If a published paper describes a particular isoform, also check if this isoform is already present in UniProt (in the **Sequences** section). You can also enter UniProt identifiers for

isoforms (a UniProtKB accession followed by a dash and a number, e.g., Q8CGC7-1). If you have an identifier from a different database, use UniProt's ID mapping service (<http://www.uniprot.org/uploadlists/>) to obtain the corresponding UniProtKB accession and retrieve the sequence.

3. The UniProt sequence retrieved is formatted to show the residue numbers, and the organism box is automatically filled in.
4. If the modification is not in the list, use the "Other" option to add it. These terms will be later mapped to the corresponding PSI-MOD terms (e.g., Ser phosphorylation will become MOD:00046). If the modification site is unknown, please enter "?" in the residue number box. Enter one modification site on each line. Use the [more] or [less] buttons to add or remove a modification line.
5. The Cytoscape view may be complex, and you may want to hide nodes to focus on a specific part of the graph. To hide all the protein nodes, select "display option" from the top menu bar and uncheck "All Protein" under "Nodes Type." Otherwise, selecting any node or set of nodes and right clicking will open a menu with the option to hide nodes.

Acknowledgments

PRO Consortium participants: Protein Information Resource, The Jackson Laboratory, Reactome, and the New York State Center of Excellence in Bioinformatics and Life Sciences. PRO is funded by NIH grant R01GM080646.

References

1. Shihab HA, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR (2014) Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics* 8:11. doi:[10.1186/1479-7364-8-11](https://doi.org/10.1186/1479-7364-8-11)
2. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446(7132):153–158. doi:[10.1038/nature05610](https://doi.org/10.1038/nature05610)
3. Omenn GS, Yocum AK, Menon R (2010) Alternative splice variants, a new class of protein cancer biomarker candidates: findings in pancreatic cancer and breast cancer with systems biology implications. *Dis Markers* 28(4):241–251. doi:[10.3233/dma-2010-0702](https://doi.org/10.3233/dma-2010-0702)
4. Menon R, Zhang Q, Zhang Y, Fermin D, Bardeesy N, DePinho RA, Lu C, Hanash SM, Omenn GS, States DJ (2009) Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Res* 69(1):300–309. doi:[10.1158/0008-5472.can-08-2145](https://doi.org/10.1158/0008-5472.can-08-2145)

5. Mair W, Muntel J, Tepper K, Tang S, Biernat J, Seeley WW, Kosik KS, Mandelkow E, Steen H, Steen JA (2016) FLEXITau: quantifying post-translational modifications of Tau protein in vitro and in human disease. *Anal Chem* 88(7):3704–3714. doi:[10.1021/acs.analchem.5b04509](https://doi.org/10.1021/acs.analchem.5b04509)
6. Kessler BM (2013) Ubiquitin - omics reveals novel networks and associations with human disease. *Curr Opin Chem Biol* 17(1):59–65. doi:[10.1016/j.cbpa.2012.12.024](https://doi.org/10.1016/j.cbpa.2012.12.024)
7. Jin K, Musso G, Vlasblom J, Jessulat M, Deineko V, Negroni J, Mosca R, Malty R, Nguyen-Tran DH, Aoki H, Minic Z, Freywald T, Phanse S, Xiang Q, Freywald A, Aloy P, Zhang Z, Babu M (2015) Yeast mitochondrial protein-protein interactions reveal diverse complexes and disease-relevant functional relationships. *J Proteome Res* 14(2):1220–1237. doi:[10.1021/pr501148q](https://doi.org/10.1021/pr501148q)
8. Climer LK, Dobretsov M, Lupashin V (2015) Defects in the COG complex and COG-related trafficking regulators affect neuronal Golgi function. *Front Neurosci* 9:405. doi:[10.3389/fnins.2015.00405](https://doi.org/10.3389/fnins.2015.00405)
9. Natale DA, Arighi CN, Blake JA, Bult CJ, Christie KR, Cowart J, D'Eustachio P, Diehl AD, Drabkin HJ, Helfer O, Huang H, Masci AM, Ren J, Roberts NV, Ross K, Ruttenberg A, Shamovsky V, Smith B, Yerramalla MS, Zhang J, AlJanahi A, Celen I, Gan C, Lv M, Schuster-Lezell E, Wu CH (2014) Protein ontology: a controlled structured network of protein entities. *Nucleic Acids Res* 42(Database issue):21
10. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25(11):1251–1255. doi:[10.1038/nbt1346](https://doi.org/10.1038/nbt1346)
11. Smith LM, Kelleher NL (2013) Proteoform: a single term describing protein complexity. *Nat Methods* 10(3):186–187. doi:[10.1038/nmeth.2369](https://doi.org/10.1038/nmeth.2369)
12. Arighi C, Shamovsky V, Masci AM, Ruttenberg A, Smith B, Natale DA, Wu C, D'Eustachio P (2015) Toll-like receptor signaling in vertebrates: testing the integration of protein, complex, and pathway data in the protein ontology framework. *PLoS One* 10(3):e0122978. doi:[10.1371/journal.pone.0122978](https://doi.org/10.1371/journal.pone.0122978)
13. Bult CJ, Drabkin HJ, Evsikov A, Natale D, Arighi C, Roberts N, Ruttenberg A, D'Eustachio P, Smith B, Blake JA, Wu C (2011) The representation of protein complexes in the Protein Ontology (PRO). *BMC Bioinformatics* 12:371. doi:[10.1186/1471-2105-12-371](https://doi.org/10.1186/1471-2105-12-371)
14. Altenhoff AM, Skunca N, Glover N, Train CM, Sueki A, Pilizota I, Gori K, Tomiczek B, Muller S, Redestig H, Gonnet GH, Dessimoz C (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res* 43(Database issue):D240–D249. doi:[10.1093/nar/gku1158](https://doi.org/10.1093/nar/gku1158)
15. Celen I, Ross KE, Arighi CN, Wu CH (2015) Bioinformatics knowledge map for analysis of beta-catenin function in cancer. *PLoS One* 10(10)
16. The Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res* 43:D1049–D1056. doi:[10.1093/nar/gku1179](https://doi.org/10.1093/nar/gku1179)
17. Montecchi-Palazzi L, Beavis R, Binz PA, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS (2008) The PSI-MOD community standard for representation of protein modification data. *Nat Biotechnol* 26(8):864–866. doi:[10.1038/nbt0808-864](https://doi.org/10.1038/nbt0808-864)
18. Mungall CJ, Batchelor C, Eilbeck K (2011) Evolution of the sequence ontology terms and relationships. *J Biomed Inform* 44(1):87–93. doi:[10.1016/j.jbi.2010.03.002](https://doi.org/10.1016/j.jbi.2010.03.002)
19. Consortium U (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42(Database issue):191–198
20. Meldal BH, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS, Gaulton A, Licata L, Melidoni AN, Ricard-Blum S, Roehert B, Skyzypek MS, Tiwari M, Velankar S, Wong ED, Hermjakob H, Orchard S (2015) The complex portal—an encyclopaedia of macromolecular complexes. *Nucleic Acids Res* 43(Database issue):D479–D484. doi:[10.1093/nar/gku975](https://doi.org/10.1093/nar/gku975)
21. Lin Y, Xiang Z, He Y (2015) Ontology-based representation and analysis of host-Brucella interactions. *J Biomed Semantics* 6:37. doi:[10.1186/s13326-015-0036-y](https://doi.org/10.1186/s13326-015-0036-y)
22. Eppig JT, Richardson JE, Kadin JA, Ringwald M, Blake JA, Bult CJ (2015) Mouse Genome Informatics (MGI): reflecting on 25 years. *Mamm Genome* 26(7–8):272–284. doi:[10.1007/s00335-015-9589-4](https://doi.org/10.1007/s00335-015-9589-4)
23. McDowall MD, Harris MA, Lock A, Rutherford K, Staines DM, Bahler J, Kersey PJ, Oliver SG, Wood V (2015) PomBase 2015: updates to the fission yeast database. *Nucleic Acids Res* 43(Database issue):D656–D661. doi:[10.1093/nar/gku1040](https://doi.org/10.1093/nar/gku1040)

24. Jensen M, Cox AP, Chaudhry N, Ng M, Sule D, Duncan W, Ray P, Weinstock-Guttman B, Smith B, Ruttenberg A, Szigeti K, Diehl AD (2013) The neurological disease ontology. *J Biomed Semantics* 4(1):42. doi:[10.1186/2041-1480-4-42](https://doi.org/10.1186/2041-1480-4-42)
25. Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, Baumgartner WA Jr, Cohen KB, Verspoor K, Blake JA, Hunter LE (2012) Concept annotation in the CRAFT corpus. *BMC Bioinformatics* 13:161. doi:[10.1186/1471-2105-13-161](https://doi.org/10.1186/1471-2105-13-161)
26. Cunningham F, Moore B, Ruiz-Schultz N, Ritchie GR, Eilbeck K (2015) Improving the Sequence Ontology terminology for genomic variant annotation. *J Biomed Semantics* 6:32. doi:[10.1186/s13326-015-0030-4](https://doi.org/10.1186/s13326-015-0030-4)
27. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 43(Database issue):27
28. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279–D285. doi:[10.1093/nar/gkv1344](https://doi.org/10.1093/nar/gkv1344)
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11): 2498–2504
30. Obsilova V, Kopecka M, Kosek D, Kacirova M, Kylarova S, Rezabkova L, Obsil T (2014) Mechanisms of the 14-3-3 protein function: regulation of protein function through conformational modulation. *Physiol Res* 63(Suppl 1): S155–S164
31. Nogueiras R, Williams LM, Dieguez C (2010) Ghrelin: new molecular pathways modulating appetite and adiposity. *Obes Facts* 3(5):285–292. doi:[10.1159/000321265](https://doi.org/10.1159/000321265)
32. Asakawa A, Inui A, Fujimiya M, Sakamaki R, Shinfuku N, Ueta Y, Meguid MM, Kasuga M (2005) Stomach regulates energy balance via acylated ghrelin and desacyl ghrelin. *Gut* 54(1):18–24. doi:[10.1136/gut.2004.038737](https://doi.org/10.1136/gut.2004.038737)

CATH-Gene3D: Generation of the Resource and Its Use in Obtaining Structural and Functional Annotations for Protein Sequences

Natalie L. Dawson, Ian Sillitoe, Jonathan G. Lees, Su Datt Lam, and Christine A. Orengo

Abstract

This chapter describes the generation of the data in the CATH-Gene3D online resource and how it can be used to study protein domains and their evolutionary relationships. Methods will be presented for: comparing protein structures, recognizing homologs, predicting domain structures within protein sequences, and subclassifying superfamilies into functionally pure families, together with a guide on using the webpages.

Key words Protein domain, Database, Sequence, Structure, Classification, Annotation, Structural model prediction, Function prediction

1 Introduction to CATH-Gene3D and Protein Structure Classifications

The first protein structure, myoglobin, was solved in the late 1950s using X-ray crystallography. As the number of solved protein structures grew, a computational archive to collect, standardize, and distribute this data was established in 1971, the Protein Data Bank (PDB) [1]. When CATH [2] was first established, there were only ~3000 protein structures in the PDB, whereas in 2016 there are now almost 110,000. This dramatic increase in size has made it impractical for users to perform an exhaustive analysis of biologically meaningful information (such as common structural units and distant evolutionary relationships) through manual inspection of individual structures. However, this additional layer of information would provide significant insight for the biological community interested in studying the structural data available in the PDB.

CATH-Gene3D is a free, publicly available online resource providing information on the evolutionary relationships of protein domains through structural, sequence, and functional annotation

data. The CATH structural classification was established in the mid-1990s and today contains 308,999 structural domains in version 4.1 (*see Note 1*). Additional sequence information for domains with no experimentally determined structures is provided by the sister resource, Gene3D. This uses the curated information in CATH to predict an additional 53 million structural domain assignments from databases of known protein sequences (v14).

The Structural Classification of Proteins (SCOP) resource [3] was also created in the mid-1990s and like CATH, provides a comprehensive protein structure classification. CATH and SCOP classify protein domains from at least 80 % of proteins in the PDB [56]. The mapping of protein folds between the two resources suggests differences in the way domains have been clustered in more general terms. However when considering the superfamily level (which groups evolutionary relatives), approximately 70 % of these clusters can be mapped between the two resources. The SUPERFAMILY resource [4] is analogous to Gene3D: a sister resource to SCOP, containing sequences from almost 2500 completely sequence genomes that have been assigned to SCOP superfamilies.

SCOP has recently been succeeded by the prototype SCOP2 classification [5]. This resource allows a number of different types of annotations and moves the structural classification from a hierarchical tree structure to a directed acyclic graph (DAG). Structural and evolutionary relationships have been separated to allow more complex relationships to be represented accurately. For example, the graph structure allows proteins that are evolutionarily related but have different folding arrangements to be linked. This was difficult in the previous hierarchical classification, in which the superfamily level was a grouping within the fold level.

Since SCOP and SCOP2 are not up-to-date with the latest version of the PDB, an extended version of SCOP, SCOPe, was recently established by the Chandonia group [6]. SCOPe uses a combination of automatic and manual curation methods to classify more recent PDB structures and also corrects some errors in SCOP. The ASTRAL database, which uses a sequence-based approach to classify recently deposited PDB protein chains into SCOPe using the structural classification hierarchy in SCOP, is also incorporated and updated. New protein chains are searched against SCOP using BLAST to look for previously classified domains that are significantly similar in sequence (i.e., have an E-value $\leq 10^{-4}$). The aligned match must also have high coverage to the query sequence—only ten residues at each end can be missing at most [6].

Another recent structural classification is the Evolutionary Classification of Protein Domains, or ECOD [7], by the Grishin group. Like SCOP, this only contains proteins with experimentally determined structures. Similar to CATH and SCOP, protein domains are hierarchically classified using evolutionary relationship

information; however, this database focuses on finding remote homologs and it is also updated every week. As of March 2016 it contained 116,441 PDB structures.

2 Materials: Algorithms Used in the CATH-Gene3D Protocol

Below we describe the algorithms used to classify structures and sequences in CATH-Gene3D. These include methods for assigning domain boundaries from structure, methods for identifying homologs methods for subclassifying families within superfamilies and detecting highly conserved sites, and methods for building 3D models of structurally uncharacterized sequences in CATH superfamilies.

2.1 Algorithms to Identify Protein Structural Domains

A protein domain can be defined as a compact, independently evolving [8] structural unit found within a protein peptide chain. Domains have a hydrophobic core, tend to fold independently, and sometimes have their own independent function [9]. One or more domains can be found in a protein chain, where their ordered combination is referred to as their multi-domain architecture (MDA). Studies of domain annotations in completed genomes have shown that over two-thirds of known proteins in prokaryotes and nearly 80 % of protein in eukaryotes contain two or more domains [8].

2.1.1 Sequence-Based Methods for Recognizing Domains in Proteins

Domain boundaries can be predicted by scanning a query protein sequence against sequence profiles or hidden Markov models (HMMs) built with SAM (Sequence Alignment and Modeling) [10] from representative domain structures, e.g., in SCOP or CATH [11]. The best matching HMM to the query is determined and if it is a significant match (i.e., it has an *E*-value less than 0.001), the model may be used to predict domain boundaries.

2.1.2 Structure-Based Methods for Recognizing Domains in Proteins

The CATHEDRAL pipeline compares a query protein structure against a library of previously classified structural domains in CATH to infer domain boundaries. High scoring matches (i.e., matches with a SSAP score above 80 and an RMSD below 6 Å) to this query protein structure can be used to predict domain boundaries. CATHEDRAL first compares the query structure against a given domain fold library using GRATH, an algorithm based on graph theory, which rapidly matches the query against putative fold matches. However, as GRATH only uses secondary structure element (SSE) information, it cannot provide the optimal alignment for all residues in a protein structure. Thus, putative matches highlighted by GRATH are realigned in a pairwise manner using the slower, but more accurate, SSAP algorithm [12]. This uses double dynamic programming to produce structural alignments.

The SSAP algorithm first calculates a view for each residue in the two proteins. The view comprises the set of vectors (i.e., a vector view) from the C_{β} atoms of a given residue to the C_{β} atoms of all other residues in the protein. SSAP compares pairs of residues between the structures that have sufficiently similar structural properties such as accessibility, secondary structure, and conformational torsion angles. The vector views for each putative equivalent pair are compared, all vectors against all, and their similarity scores stored in a 2D path matrix. The first round of dynamic programming is then used to find the highest scoring path through this matrix giving the best alignment of the vector views. Scores along this highest scoring path are then added to a summary score matrix. Once all putative equivalent residue pairs have been compared in this way, dynamic programming is used again to find the optimal path through this summary matrix giving the final global alignment of the two protein structures [13].

2.1.3 *Ab initio*-Based Methods for Detecting Domain Boundaries in Protein Structures

DETECTIVE [14], DOMAK [15], and PUU [16] are three *ab initio* methods designed to detect one or more protein domains within a given protein structure. Each algorithm defines putative domain boundaries using a different method, though each uses atomic interaction information. DETECTIVE searches for the hydrophobic core of each domain and residues in contact with this core, DOMAK randomly splits a protein into two and analyzes the number of atomic contacts between and within each half to determine if the split could represent different domains, and PUU calculates the oscillation of residue clusters as a single domain has its own pattern of motion relative to other domains in the same protein. All three methods are able to identify discontinuous domains and are discussed in more detail below.

The DETECTIVE algorithm is based upon the idea that a domain has a hydrophobic core. An independent core is found using secondary structure, side-chain accessibility, and side-chain-side-chain contacts information. More specifically, the core is identified by finding nonpolar residues with low solvent accessibilities that are within regular secondary structures (i.e., an alpha helix or a beta-strand), and that interact with other nonpolar residues [17]. When the core has been defined, this central part of the domain is built upon to initiate the assignment of domain boundaries. The remaining residues in the protein are each searched against the core residues, and if atomic contacts are made between a given pair of residues (i.e., when two atoms are closer together than the sum of their van der Waals (vdWs) radii plus 1 Å [17]), the residue is provisionally assigned to that domain. Isolated residues are removed and finally, if there are unassigned residues remaining, the domain ends are extended to the ends of secondary structure elements, and then extended to the N- and C-termini [14].

The DOMAK algorithm is built upon the Rossmann and Liljas concept [18] that protein domains have more internal contacts than external. A protein is split into two arbitrary parts (A and B); the number of internal contacts (residues within 5 Å) is counted within each part and the number of external contacts between A and B is calculated. As domains are compact entities, the ratio of internal to external contacts will be large for a protein when the two parts are distinct and therefore belong to different domains. This process is iterated until all sets of arbitrary parts have been analyzed and the best ratio is found [15].

PUU is based upon the concept that protein domains are compact entities that interact through noncovalent atomic interactions. Each domain in a multidomain protein has its own continuous motion in relation to neighboring domains, which is defined by the strength of the atomic interactions at the domain's interface and the distribution of the domain's mass. A time constant of relative motion (t) is calculated between clusters of residues by counting atoms and their contacting atom pairs. The higher the value of t , the more likely it is that the clusters of residues belong to different domains [16].

While each of these three methods reported 70–80 % accuracy when they were benchmarked using a manually curated dataset from CATH, they frequently (~80–90 % of the time) differed in their boundary assignments suggesting that, although they provide a good initial guide to domain boundaries, some level of manual curation is advisable [19].

2.2 Algorithms Used to Recognize Protein Homologs

Protein 3D structure is generally more highly conserved than protein sequence, with the exception of some homologs whose structures are highly diverged. These very remote homologs can still be recognized if they contain certain highly conserved sequence motifs, associated with the domain [20]. Furthermore, structural similarity alone is not a sufficient indicator of homology as there are clearly constraints on the ways in which proteins fold and secondary structures pack together. Therefore, it is advisable to use both structure- and sequence-based methods in the search for homologs.

2.2.1 Structure-Based Comparison Methods

The CATHEDRAL structure comparison pipeline [21], described in Subheading 2.1 above, is used to find the most significant fold matches in CATH for a query protein.

Other popular structural comparison algorithms that are also used in the CATH classification protocol to validate structural similarities include: STRUCTAL [22] and TM-align [23]. STRUCTAL starts with an initial alignment based on a preliminary sequence alignment of the two structures. It then superimposes the structures, infers a new structure-based alignment from the superposition, and iterates. It uses dynamic programming and relies on a Kabsch RMSD

rotation matrix to find the optimal rotation and translation in terms of RMSD between the structures being aligned [24].

TM-align performs a pairwise protein structural alignment using three different methods to generate initial structural alignments, which are then superposed using the TM-score rotation matrix derived from these initial alignments. The three different methods involve: (1) the use of dynamic programming to structurally align secondary structures, (2) gapless threading of the smaller structure onto the larger structure, and (3) the use of dynamic programming on a summary 2D matrix that combines the scoring matrices from steps (1) and (2). The process of creating alignments and superpositions is iterative and repeated until the alignment is stable [23].

The TM-score (i.e., Template Model-score) calculated by TM-align ranges from 0 to 1.0, and was originally developed to measure structural similarity between a protein model and an experimentally solved structure, but it can also be applied to pairs of experimentally solved structures. A score greater than 0.5 generally reflects that the two protein structures have the same fold in CATH or SCOP, for example [23].

2.2.2 Sequence-Based Comparison Methods

Pairwise sequence comparison methods are used to identify closely related homologs and are much faster than most residue-based structure comparison methods as they are simply comparing linear strings. Pairwise Needleman-Wunsch-based sequence alignments can be used to find closely related homologs i.e., domains sharing at least 35 % sequence identity.

Profile hidden Markov models (HMMs) exploit residue preference information in a multiple sequence alignment of protein relatives and are similar to sequence profiles in that they calculate amino acid frequency at each alignment position. In addition, they also calculate the probabilities for insertions and deletions at each position [25], providing a probability distribution over a significant number of diverse sequences [26]. The jackhmmmer algorithm from the HMMER3 suite [27] uses an iterative process to search a query protein sequence against a large protein sequence database (e.g., UniProt [28]) to create an increasingly sensitive profile HMM.

The Profile Comparer (PRC) algorithm [29] performs a pairwise HMM comparison by aligning a profile HMM against a library of profile HMMs and calculates a score. An E-value is reported to indicate the significance of the score.

Another popular tool for detecting remote protein homologs is HHpred [30], which can also be used in structure prediction to create alignments between a query sequence and suitable templates. The HHpred webserver searches a query sequence or a multiple sequence alignment against a wide choice of databases such as PDB, SCOP [3], and Pfam [31]. For a given query

sequence, an alignment is built using PSI-BLAST [32] then a profile HMM generated from this alignment. The query HMM is then compared with each HMM in the selected database using the HHsearch software [30].

2.3 Algorithms to Subclassify Superfamilies

The FunFHMMER algorithm is a sequence-based method used to subclassify homologous superfamily members into functional families, or FunFams. FunFams ideally represent groups of sequences that perform the same or very similar functions. For each superfamily, members are grouped using a profile-based, hierarchical agglomerative clustering method [33]: all sequences in the superfamily are first clustered into groups with at least 90 % sequence identity. Sequences in these groups are multiply aligned to generate sequence profiles, which are then compared in a pairwise manner and merged if they display significant similarity. This process is iterated to construct a hierarchical tree reflecting sequence relationships within the superfamily.

Subsequently, functionally coherent families or FunFams are identified using a method that decides where to cut the tree, i.e., whether two nodes (i.e., sequence clusters) can be considered functionally related. This exploits the GroupSim method of Capra and Singh [34] to recognize differences in specificity-determining residues (SDRs) between sequence clusters associated with the nodes. These are residues that are differentially conserved in the two clusters being compared.

2.4 Algorithms to Identify Conserved Residue Sites

The Scorecons algorithm [35] is used to find highly conserved sites within CATH functional families that could coincide with functional sites. This method uses entropy information to quantify the amino acid conservation at each residue position in a multiple sequence alignment. A conservation score between 0 and 1 is assigned to each alignment position, where 0 reflects no conservation and 1 reflects complete conservation. A Dayhoff-like data matrix [36] is used to calculate amino acid diversity at each alignment position. The total diversity of the positions in the alignment is also calculated as the DOPs score (Diversity Of Positions), which takes into account the number of different conservation scores and their relative frequency. This score ranges from 0 to 100, where 0 indicates no diversity and that all positions have the same conservation score, through to 100 where all positions have different scores [35]. Multiple alignments having a DOPS score of 80 or more are considered sufficiently diverse to identify conserved sites.

2.5 Algorithms to Find Structural Templates and Build Structural Models

Structural models are built for structurally uncharacterized sequences in CATH FunFams where there is at least one structural relative available to use as a template structure. Sequences within each FunFam are first aligned using the HHalign algorithm [25] and this alignment is then submitted to the MODELLER homology modeling platform

developed by the Sali Group [37, 38]. Model quality is assessed using the DOPE (Discrete Optimized Protein Energy) algorithm provided by the MODELLER suite. DOPE is a statistical potential based on the finite size and spherical shape of proteins. Medium and high quality models (i.e., models with a normalized DOPE score less than zero) are then displayed on the Gene3D website.

3 Methods: The CATH-Gene3D Platform

3.1 Identifying New Protein Domain Structures

New protein entries are downloaded every week from the PDB and split into constituent polypeptide chains, which are processed separately (*see* a simplified flow chart in Fig. 1). A fast sequence scan is performed against all other chains waiting to be processed: those chains with at least 35 % sequence identity are retained in a “holding pen” so as to reduce the redundancy and volume of the structures being processed. Those chains with less than 35 % sequence identity are compared with other chains that have already been “chopped” into domains. If there is a previously annotated chain with high sequence similarity and residue overlap ($\geq 80\%$ sequence identity, 80 % overlap of residues in the larger chain with smaller chain), the domain predictions are inherited onto this new chain.

Otherwise, additional sequence- (CATH-HMM-Scan), structure- (CATHEDRAL [21]), and *ab initio*-based (PUU [16]),

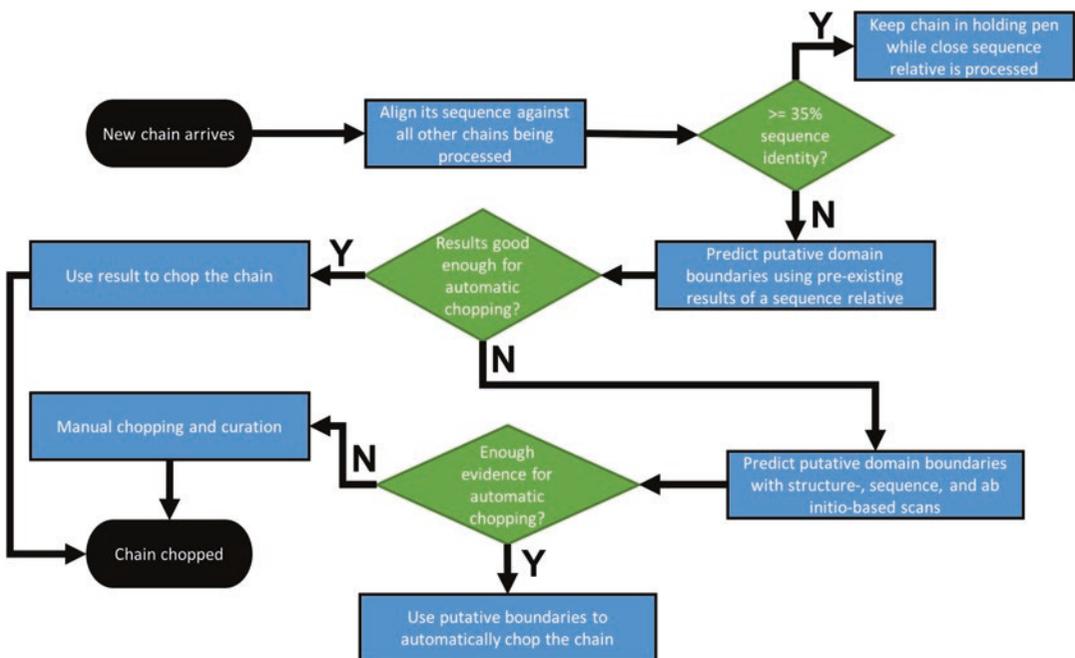


Fig. 1 Steps taken to predict putative domain boundaries and use them to “chop” a protein chain into domains, either automatically or through manual curation

DETECTIVE [14], DOMAK [15]) scans are performed to provide predictions of the most likely structural domain boundaries. In cases where a confident automated prediction cannot be made, human experts assign domain boundaries manually (through internal web pages).

3.2 Assigning Domains to Superfamilies

The structures and sequences of newly identified structural domains are then compared against a library of nonredundant domains previously classified in CATH superfamilies to search for domains sharing sufficient levels of similarity to infer that the domains share a common evolutionary ancestor.

For classifying batches of new domains, domains are first clustered into homologous groups to reduce the time and effort required for classification. The new domains are compared sequence-wise against all other domains to be processed: again, those domains with at least 35 % sequence identity are temporarily retained in a “holding pen” (as previously described). Domains with less than 35 % sequence identity (i.e., those novel enough to act as sequence representatives) are compared against domains assigned in CATH using HMM-HMM (PRC) and structure-based (CATHEDRAL) comparison methods. The HMM-HMM-based algorithms have been carefully benchmarked and are often sensitive enough to detect remote homologs [19].

HMM libraries are produced for CATH superfamilies using HMMER3 software [27]. Multiple HMMs are built for each superfamily; one for each cluster of superfamily members grouped at 35 % sequence identity (i.e., for each superfamily S35 group; *see* Subheading 3.3 for details). These models are built by using the HMM building tool jackhammer [27] to search against the UniRef90 sequence database [28] with five iterations. This ensures that more remote homologs are incorporated in the HMM, which in turn increases the sensitivity of searches against the resulting HMM. A pairwise comparison of these HMMs is performed using the PRC algorithm [29], which aligns a profile HMM of the query against a library of profile HMMs.

The structures of newly identified domains are scanned against a library of nonredundant structural representatives from each CATH superfamily using the CATHEDRAL pipeline described in Subheading 2.1.2.

Two out of three of the following criteria need to be met for domains to be considered homologous: significant sequence similarity (≥ 35 % or a statistically significant E-value from the PRC comparison), high structural similarity (SSAP score ≥ 80), evidence in the literature, or a public repository of experimentally supported functional similarity. If the information obtained for a new query domain satisfies these criteria, the domain can be automatically assigned to the same CATH superfamily as the matched proteins; otherwise, manual curation by human experts is performed (*see* **Note 2**).

Structural and sequence matches to several diverse relatives within a CATH superfamily provide added confidence to the homology assignment.

The detection of homology can be difficult in the case of very distantly related (i.e., remote) homologs. A supervised support vector machine (SVM) algorithm has recently been developed which combines all the available evidence on structural and sequence similarity for known relatives in CATH and provides a powerful predictor capable of identifying remote evolutionary relationships. CATH-SVM was trained by providing a list of homologous and nonhomologous domain pairs as examples of positive and negative results. All pairs were associated with scores of structure (SSAP) and sequence (PRC) similarity along with a number of other metrics (size, residue overlap, etc.). The SVM then learns the optimal way of combining these scores to distinguish whether or not two domains are homologous. Careful benchmarking using a set of manually curated homologs allows more than 85 % of new domains to be assigned to an existing superfamily in CATH (with an error rate of 5 %).

3.3 The CATH Classification of Protein Structures: the Hierarchy

Domains are assigned to different levels in the CATH hierarchy using a variety of information, which is dependent upon the level. There are four major levels: class (C), architecture (A), topology or fold group (T), and homologous superfamily (H) (*see* Fig. 2).

At the Class or C-Level, domains are clustered according to the percentage of α -helices or beta-strands that they contain. This level is split into four categories: mainly alpha, mainly beta, mixed alpha and beta, and few secondary structures. Domains are then grouped into their architectures at the A-Level, which is determined by how the secondary structures are arranged in 3D space. There are 40 distinct architectures in CATH version 4.1. Assignments to the C and A levels are mostly manual. The top three most highly populated architectures (in terms of structural domain count) in the three main structural classes are (in descending order for each class): Orthogonal Bundle (1.10), Up-down Bundle (1.20), Alpha Horseshoe (1.25); Beta 2-layer Sandwich (2.60), Beta Barrel (2.40), Beta Roll (2.30); Alpha-Beta 3-Layer(aba) Sandwich (3.40), Alpha-Beta 2-Layer Sandwich (3.30), Alpha-Beta Barrel (3.20). These architectures comprise approximately 81 % of all structural domains in CATH.

The T-Level provides information on how secondary structure elements are connected and their 3D arrangement. Domains are clustered at the T-level according to their general structural similarity. At the H-Level, domains are only grouped if there is good evidence that they are related by evolution, i.e., they are homologous (*see* Subheading 3.2 for more details).

In addition to the major levels of the hierarchy there are further levels representing different degrees of sequence similarity.

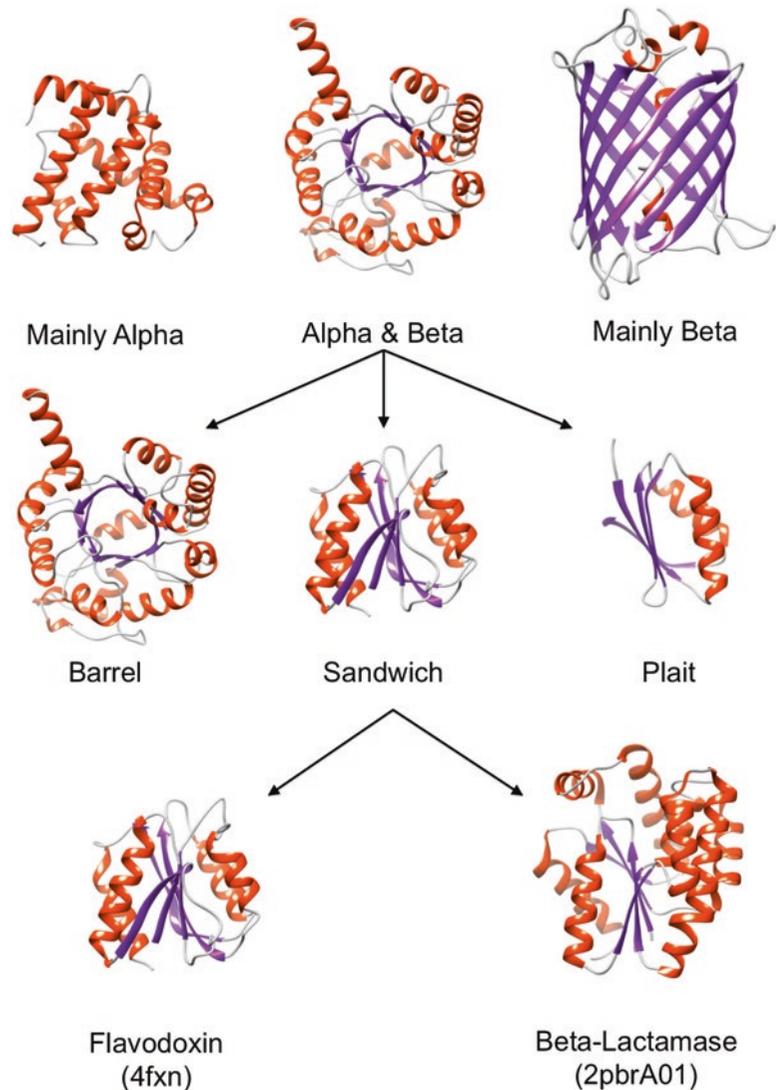


Fig. 2 The top three levels of the CATH structural domain classification hierarchy: Class (or C-level) that is based on the secondary structure content, Architecture (A-level) that captures how the secondary structures are arranged in three-dimensional space, and Topology or fold (T-level) that provides information on the connectivity of the secondary structure elements and their three-dimensional arrangement

For example, within each H-Level, domains are grouped using multi-linkage clustering at different significant sequence similarity levels: 35, 60, 95, and 100 %. A list of representative domains is provided for each sequence cluster within each CATH superfamily (e.g., S35, S60 clusters), with each representative symbolizing a group of domains with a common minimum sequence identity. For example, within a homologous superfamily there may be one

or more S35 clusters and each domain within a S35 cluster shares at least 35 % sequence identity. More recently introduced to CATH, a nonredundant data set analogous to the ASTRAL40 data set (based on the SCOP classification) [54, 55] is also provided.

3.4 Subclassification of Superfamilies

Some superfamilies in CATH-Gene3D can be very large and diverse in terms of sequence, structure, and function. Figure 3 shows that the largest 100 superfamilies (in terms of the number of structural domains within them) contain over half the total number of structural domains in CATH.

As previously mentioned, superfamily members are subclassified in terms of sequence identity to produce the SOLID sequence clusters (i.e., relatives clustered at 35 % (S), 60 % (O), 95 % (L), and 100 % (I) sequence identity respectively. The final cluster (D) contains the individual domains within the superfamily).

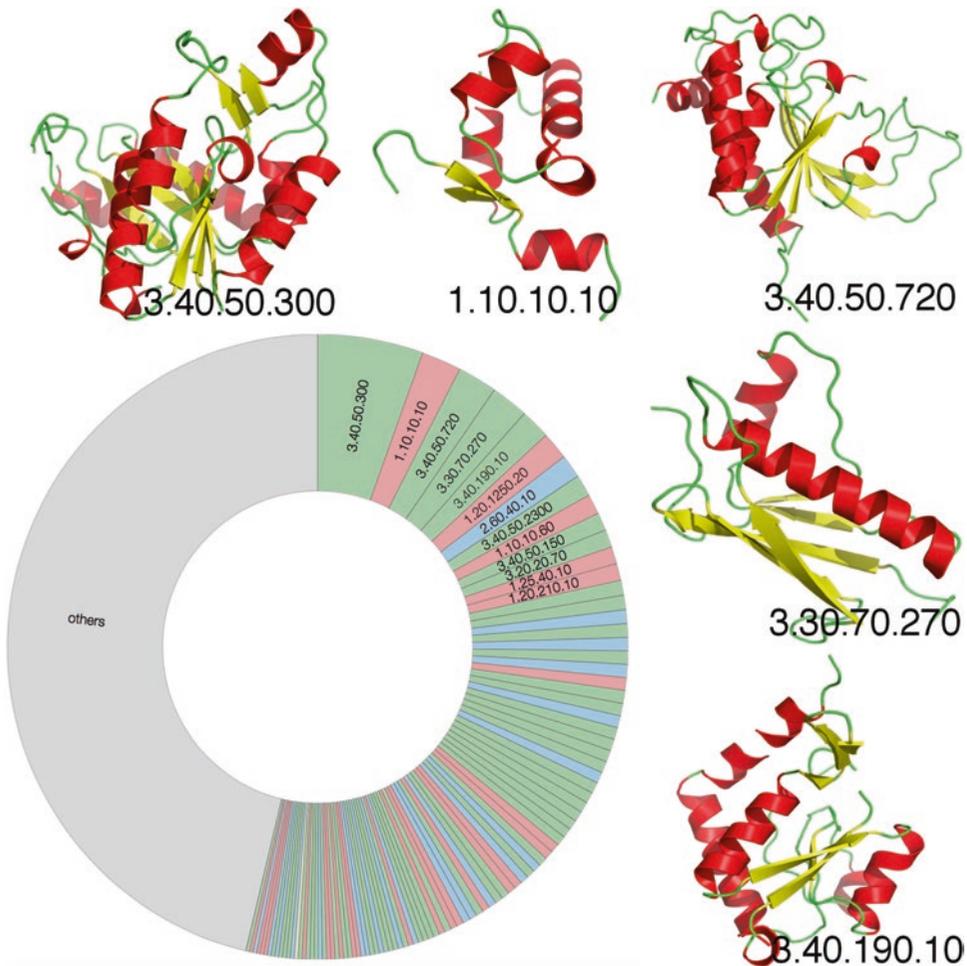


Fig. 3 The most populated 100 CATH superfamilies according to the number of predicted domains. An example nonredundant sequence representative from each of the top five superfamilies is shown

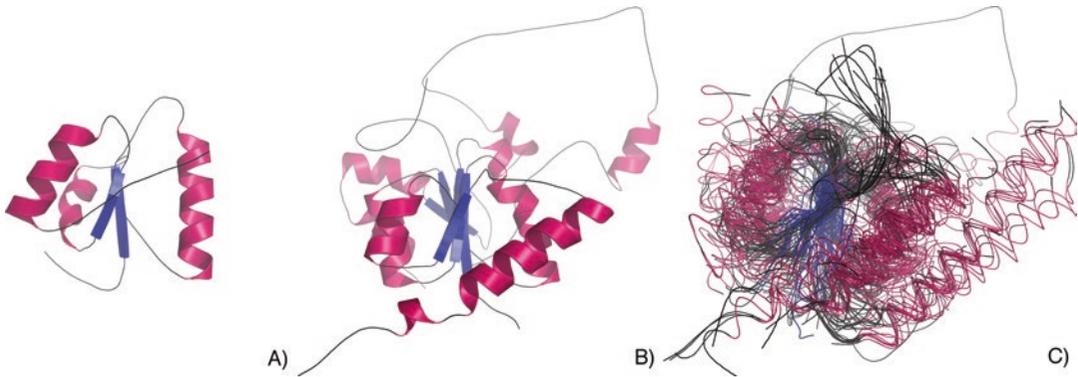


Fig. 4 Structural diversity in the “Nitrogenase molybdenum iron protein domain” superfamily (ID: 3.40.50.1980). This is a highly structural diverse superfamily with eight SSG9 clusters. The smallest (**a**) and largest (**b**) domain structures from the superfamily, and a superposition of all nonredundant relatives (at 35 % sequence identity) in the superfamily (**c**)

Structurally similar groups, or SSGs, are subgroups of superfamily members where the structures can be superposed with an RMSD less than 9 Å. Even in a superfamily with highly diverse sequences and structures (*see* the example in Fig. 4), the superposition in Fig. 4c shows that the related protein structures can remain highly conserved, as shown by the visible structural “core.”

3.4.1 Subclassification into Functional Families

The FunFHMME algorithm (*see* Subheading 2.3) is used to subclassify CATH superfamilies into FunFams. Independent validation of this approach by the international experiment, CAFA (Critical Assessment of Function Annotation) has shown FunFams to be highly competitive in the assignment of functional annotations to uncharacterized sequences [39]. FunFams have also been shown to be useful for repurposing drugs [40] and for selecting suitable structural templates when building 3D models. There are almost 93,000 FunFams in the latest version of CATH (v4.1), which comprise 42 % of all domain sequences in CATH-Gene3D.

3.5 The Addition of Sequence Relatives and Structural Models

There has been an exponential increase in sequence data in recent years and only a small percentage of these sequences have experimentally characterized 3D structures. As of February 2016 there are, for example, over 60 million protein sequences in the UniProt Knowledgebase (UniProtKB) (<http://www.uniprot.org>), whereas there are ~110,000 protein structures in the PDB (<http://www.ebi.ac.uk/pdbe/>). However, we know that protein domains that share significant sequence similarity adopt very similar structures, so we can use fast sequence searching tools to predict the locations of CATH structural domains for these genomic sequences. Additionally, these protein sequences have a great deal of important and useful experimental annotation, which, once associated with CATH domains, can be brought back into our database.

This has allowed us to improve our understanding of the sequence/structure/function relationships within superfamilies.

Protein domain sequences that lack experimental structural evidence are stored in the Gene3D resource. Gene3D is principally a resource of protein domain assignments using profile Hidden Markov Models (pHMMs) [41] built using known CATH domains as the starting sequence. Briefly, an HMM is built for each of the CATH S95 representative superfamily domain sequences using jackhmmer (*see* Subheading 2.2.2), with UniRef90 as the reference sequence database. Inclusion bit-score cut-off values for each HMM model are determined by benchmarking the set of known CATH domain assignments from the PDB (using a residue-based mapping to UniProt via the SIFTS resource [42]). Since the FunFams have been shown to be structurally coherent groupings of relatives [43] and therefore likely to provide good domain boundaries in new sequence relatives mapped to the FunFam, we also build HMMs for each FunFam from the multiple sequence alignments of relatives in the FunFam. A combined library of these HMMs is then used to predict domain assignments in query sequences using hmmsrch from the HMMER3 software package.

Once each domain has a predicted superfamily assignment, the set of FunFam HMMs in these superfamilies are scanned to provide functional annotations and potential refinements to the domain boundaries.

The domain assignment process can result in overlapping assignments from different HMMs and we have been using a modified version of the maximum weighted clique algorithm, DomainFinder [44] to resolve these overlaps. However, for some sequences this method fails to run (e.g., Titin), even on machines with large amounts of memory. To address this issue and enable domain assignments from a mixture of different HMM libraries, a new algorithm was implemented. This uses dynamic programming to find the optimum domain assignment using an adaptation of the Weighted Interval Scheduling algorithm for DOMain assignment (WISdom). This method picks the set of nonoverlapping domain assignments with the highest sum of bit-scores, with some adaptations from the basic weighted interval scheduling to allow a small level of overlap. These allowed overlaps are subsequently trimmed back to different degrees depending on the confidence of the prediction.

This algorithm has proven to be both fast and portable and allows the Gene3D assignment pipeline to be provided to users as a standalone package that can be run locally, for example on metagenomics datasets. The current release of Gene3D (v14.0) assigns nearly 54 million CATH domains to over 43 million protein sequences for ~20,000 cellular genomes [45]. The number of domain assignments extends to over 65 million domain

assignments when Pfam domains are added to regions not assigned a CATH domain.

3D structural models have recently been added to the Gene3D resource to increase the amount of structural information available. Over 63,000 structural models have been generated for structurally uncharacterized sequences in the Human and *Drosophila melanogaster* genomes. These models were built using FunMod, an in-house structural modeling platform. FunMod identifies structural templates for query sequences using FunFHMMer [46], BLAST [47], and HHsearch [25]. FunFHMMer is used to recognize domains that can be assigned to a CATH FunFam. Such domains are aligned against the FunFam multiple alignment using HHalign. Otherwise, to model sequences unrelated to domains in CATH FunFams, sequences are scanned against PDB sequences using BLAST and HHsearch to identify significant matches (determined using safe E-value thresholds). The sequence alignment tool HHalign [25] is used to align the query and template sequences before submitting them to the comparative modeling program MODELLER [38]. FunMod generated models based on FunFam selected templates and alignments have been shown to be of higher quality than those generated by selecting templates using default BLAST or HHpred search strategies and therefore FunFams are always searched first [45].

The Gene3D data can be used in a number of ways. Details for an individual protein's domain assignments can be quickly accessed using the search box on the Gene3D website (<http://gene3d.biochem.ucl.ac.uk/>). The data can also be used in a comparative genomics setting, i.e., to compare domain assignments between species and analyze lineage-specific innovations [48] (*see* Fig. 5 for an example of how superfamilies can expand in different ways in different organisms).

Information from other resources has also been integrated in Gene3D, for example protein annotations from UniProt [28]. These include genetic variation and post-translational modification (PTM) data. Biologists can use this data to examine how residue mutations are distributed in a particular domain structure of a 3D model.

Gene3D also imports drug annotations from DrugBank [49] and interaction data from the curated IntAct Molecular Interaction database [50]. For many interactions in IntAct there is information on subregions of the proteins that participate or influence the interaction. Many of these subregions overlap well with domain regions in Gene3D and we provide a domain-centric means of visualizing these data in Gene3D (Fig. 6). The data for Gene3D is made available for all Ensembl Genomes on the web site (http://download.cathdb.info/gene3d/CURRENT_RELEASE/).

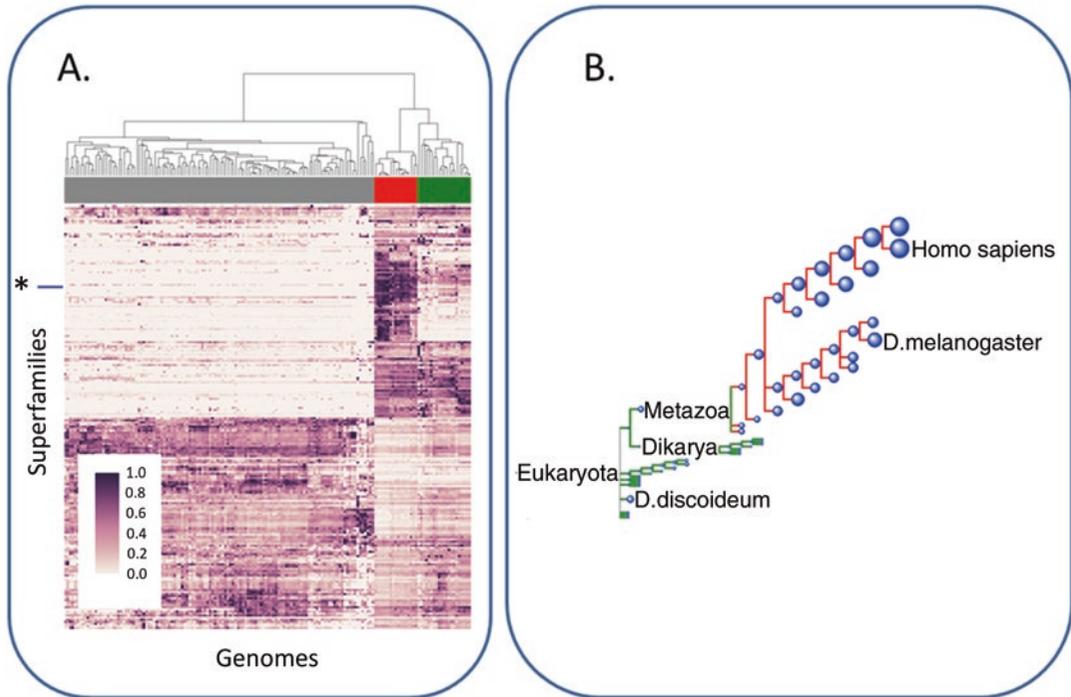


Fig. 5 Comparative genomic analysis of the top 200 CATH superfamilies across the pan-compara genomes. (a) Shows a normalized heatmap where the depth of color corresponds to a greater number of FunFams for that superfamily. Columns correspond to genomes (*red* = Metazoans, *green* = Eukaryotes, *gray* = Prokaryotes). Rows correspond to superfamilies: the row marked with an *asterisk* and *blue bar* is for the SHC-adaptor CATH superfamily (CATH ID: 3.30.505.10), which has a function in phospho-tyrosine binding. From the heatmap it can be seen this superfamily clusters with superfamilies that have expanded (in terms of FunFams) in Metazoans. (b) NCBI taxonomic tree of pan-compara eukaryotic genomes, with the node size corresponding to the number of different FunFams for the SHC-adaptor superfamily

4 Using the CATH-Gene3D Resource

This section provides step-by-step descriptions on how to access information in the CATH-Gene3D database using: the classification hierarchy (Subheading 4.1), the homologous superfamily pages (Subheading 4.2), a new protein structure (Subheading 4.3), or anew protein sequence (Subheading 4.4). Accession of the FunFam data is described in Subheading 4.5. The starting point for all this information is the CATH-Gene3D home page at www.cathdb.info (see Fig. 7).

4.1 Browsing the Classification Hierarchy

Browsing the CATH classification hierarchy provides access to all the domains classified in CATH, their evolutionary relationships, and their structural properties.

1. Select the red “Browse” button on the CATH-Gene3D homepage (see Fig. 8a).

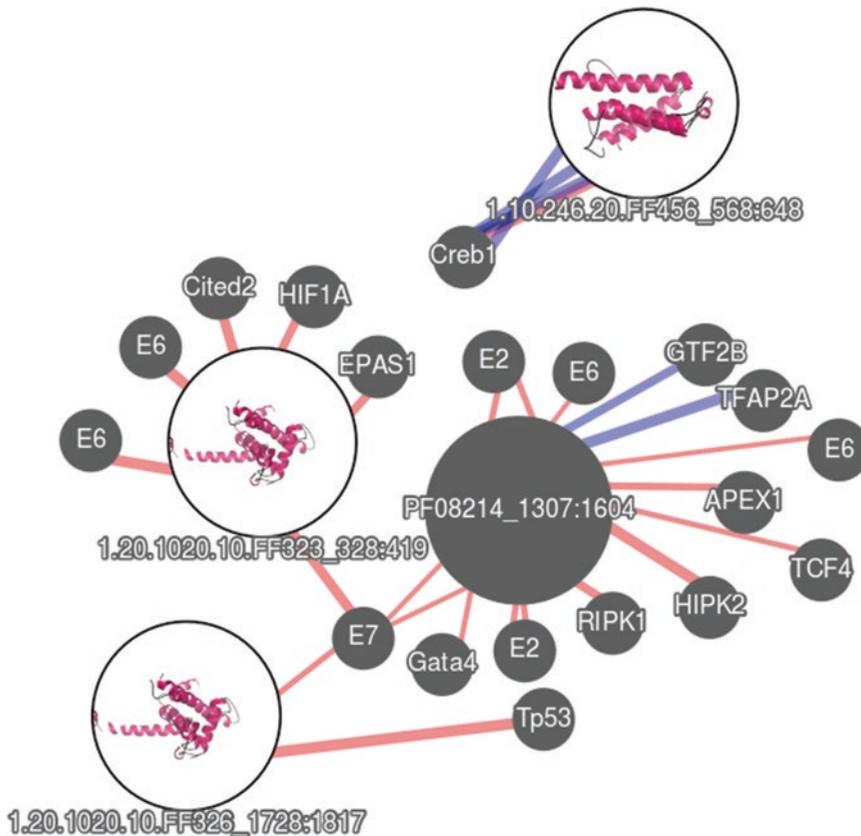


Fig. 6 Example subregion interactions for a selection of P300 protein domains. The large circular nodes (with images in the case of available CATH domains) show example domains from P300_HUMAN and the label shows its domain family code and amino acid regions, while the smaller nodes represent proteins that have evidence of interacting with that region. The width of the edge indicates how specific the annotation is to the sequence covered by the domain. *Blue links* indicate where mutations are known to affect the interaction in that domain

2. The CATH structural classification hierarchy can be viewed with either the tree browser or the sunburst diagram.
3. The tree browser provides access to the hierarchy as though it is a file system. At the root of the tree there are the four main CATH classes. Click on an entry to expand a class “C” level (e.g., 1 Mainly Alpha; *see* Fig. 8b) and reveal the architecture “A” levels below.
4. Information on the left-hand side of the page will appear throughout browsing the hierarchy, detailing the number of domains classified at the current selected level and at each subsequent hierarchical level, together with an example domain for the selected level of the hierarchy (i.e., the class level; *see* Fig. 8c). This example domain can be selected to find out more information. The PDB file for the domain can be downloaded by clicking on the link named, “[PDB].”

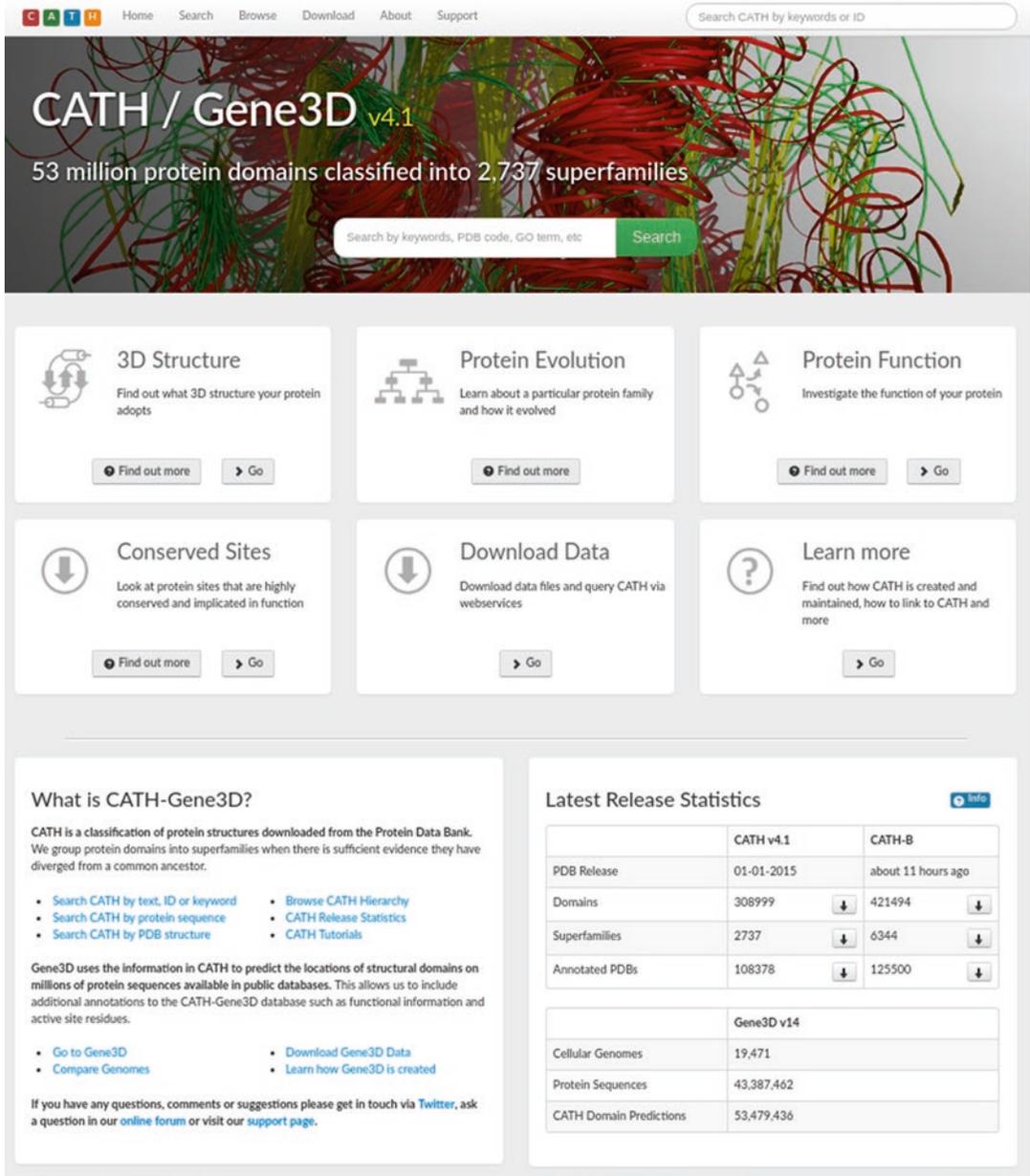


Fig. 7 The CATH-Gene3D home page, providing links to browse the CATH hierarchy, and search CATH by sequence (FASTA) or structure (PDB). There is also a free text search bar at the top right of the page

5. Click on an entry to expand an A level (e.g., 1.25 Alpha Horseshoe) and reveal the topology/fold “T” levels below.
6. Click on an entry to expand a T level (e.g., 1.25.10 Leucine-rich Repeat Variant) and reveal the homologous superfamilies “H” within.
7. Click on the superfamily entry you wish to explore further (e.g., 1.25.10.10 Leucine-rich Repeat Variant). A green button,

A)

CATH / Gene3D v4.1
53 million protein domains classified into 2,737 superfamilies

Search by keywords, PDB code, GO term, etc Search

B)

Browse CATH-Gene3D Hierarchy

BROWSE LINKS
Browse Hierarchy
Highly Diverse Superfamilies
Superfamily Comparison

Select a CATH node...

Tree Sunburst

Top of CATH Hierarchy (4 Classes)

- 1 Mainly Alpha
- 2 Mainly Beta
- 3 Alpha Beta
- 4 Few Secondary Structures

5 Architectures, 396 Folds, 908 Superfamilies, 61579 Domains
20 Architectures, 241 Folds, 547 Superfamilies, 78049 Domains
14 Architectures, 628 Folds, 1160 Superfamilies, 165745 Domains
1 Architectures, 108 Folds, 122 Superfamilies, 3626 Domains

C)

Browse CATH-Gene3D Hierarchy

BROWSE LINKS
Browse Hierarchy
Highly Diverse Superfamilies
Superfamily Comparison

Select a CATH node...

Tree Sunburst

Top of CATH Hierarchy (4 Classes)

- 1 Mainly Alpha
 - 1.10 Orthogonal Bundle
 - 1.20 Up-down Bundle
 - 1.25 Alpha Horseshoe
 - 1.40 Alpha solenoid
 - 1.50 Alpha/alpha barrel
- 2 Mainly Beta
- 3 Alpha Beta
- 4 Few Secondary Structures

5 Architectures, 396 Folds, 908 Superfamilies, 61579 Domains
287 Folds, 545 Superfamilies, 43441 Domains
101 Folds, 330 Superfamilies, 15652 Domains
5 Folds, 27 Superfamilies, 1714 Domains
1 Folds, 1 Superfamilies, 11 Domains
2 Folds, 5 Superfamilies, 761 Domains
20 Architectures, 241 Folds, 547 Superfamilies, 78049 Domains
14 Architectures, 628 Folds, 1160 Superfamilies, 165745 Domains
1 Architectures, 108 Folds, 122 Superfamilies, 3626 Domains

Select a CATH node...

C Mainly Alpha 1

CATH ID	1
Architectures	5
Topologies	396
Superfamilies	908
Domains	61579
Example Domain	1xmkA00 [PDB]



Fig. 8 Browsing the CATH structural classification hierarchy. **(a)** The classification can be accessed by clicking on the red “Browse” button. **(b)** Subsequent levels of the classification can be explored by selecting a class. **(c)** The number of different instances at each subsequent classification level is provided on the left-hand side (see left-hand red box), and all instances of the architecture, “A,” level are displayed

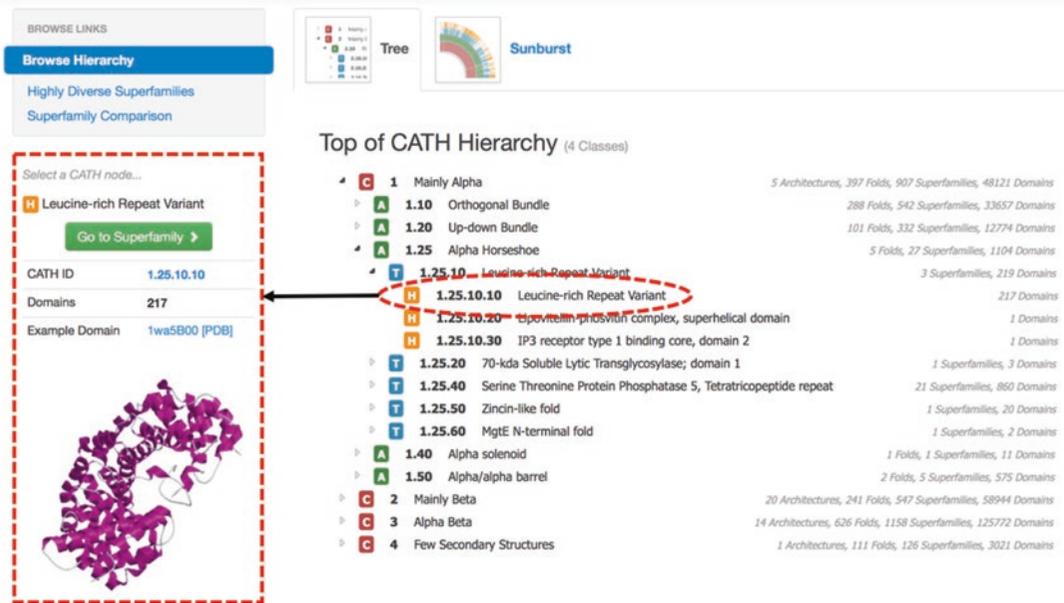


Fig. 9 Example of the CATH hierarchical structure for domain classification. Clicking on a superfamily entry in the tree browser brings up summary information, including the number of domain members, and an example domain ID and image (*left-hand red box*). A green button on the left-hand side, “Go to Superfamily,” is provided that opens up the corresponding superfamily home page when selected

“Go to Superfamily,” will appear on the left-hand side (*see Fig. 9*). Select this button to be taken to the domain superfamily page.

8. To view the CATH hierarchy with the sunburst diagram, click the “Sunburst” tab toward the top of the page.
9. The root of the tree is represented by the innermost circle (in gray). Moving outward, the red layer represents the four classes, which are each split into their architectures (in green), topologies (in blue), and finally homologous superfamilies (in yellow). Move the mouse of each section to view its name, ID, and representative domain.
10. Click on any section to redraw the diagram and only show sections within that branch of the tree. To restore the diagram to its original form, click on the gray innermost circle.

4.2 Exploring the Superfamily Pages

As of 2016, almost 308,999 domains have been assigned to 2737 superfamilies. Figure 10 shows the summary web page for an example superfamily, through which all the domain superfamily relatives can be explored.

1. Enter a domain superfamily ID into the search bar of the CATH-Gene3D home page (e.g., 3.40.50.620 for the HUPs superfamily).
2. Select the relevant superfamily from the search results.

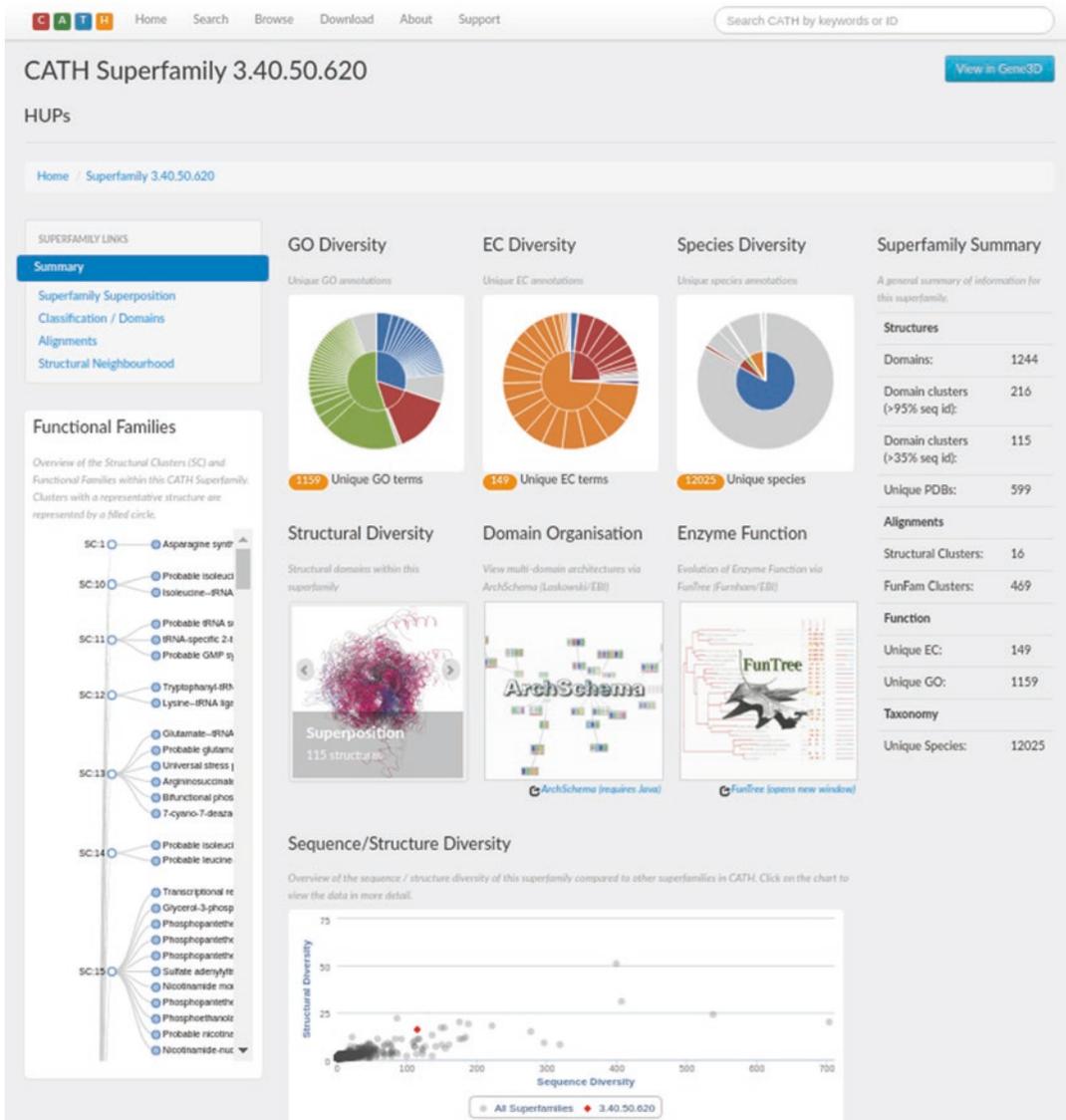


Fig. 10 Example of the summary web page provided for each superfamily. Users can access information on GO terms, EC, and species diversity, together with structural diversity, domain organization (through ArchSchema), and enzyme function diversity (through FunTree). The Sequence/Structure Diversity plot at the bottom of the page indicates the relative position of the selected superfamily (*red dot*) in terms of its members' sequence diversity (i.e., number of S35 clusters, *see* Subheading 3.3 for details) and structure diversity (i.e., number of SSGs, *see* Subheading 3.4 for details) compared to all of CATH (*gray dots*). On the left-hand side of the page, all functional families having at least one experimental GO term characterization are listed in a tree, clustered into structural similarity groups. Users can click on these to view functional family information (*see* Subheading 4.5)

3. The domain superfamily web page provides many different types of information (*see* Fig. 10). A superfamily summary is provided on the right-hand side detailing the sequence, structure, function, and species diversity.

4. To explore the sequence diversity, an overview of the structural domains within a superfamily is provided under “Classification/ Domains” in the left-hand menu, which shows how the domains group into sequence-based similarity clusters (*see* Subheading 3.4 for more details on sequence similarity clusters). At the bottom of the superfamily summary page, a comparison of sequence diversity (based upon the number of S35 sequence families) against structural diversity for all superfamilies is plotted. Your query superfamily is shown as a red dot, with all other superfamilies in gray to put it into context in terms of sequences and structure diversity.
5. To explore the structural diversity, look at the “Superfamily Superposition” and the “Structural Neighborhood” pages using the left-hand menu on the superfamily summary page. The superfamily superposition is an effective way of observing the structural conservation and diversity across its members. S35 representative domains are used to create the superposition, which can be downloaded as an image or a PyMOL file. The structural neighborhood page describes structural neighbors inside and outside the superfamily. An interactive plot is loaded for both, where clicking on a point will load a window describing the pair of structural domain neighbors in terms of their superfamily IDs, length, sequence identity, sequence overlap, normalized RMSD, and SSAP score.
6. To explore the functional diversity, mouse-over the “GO Diversity” and “EC diversity” pie charts, click on their “Unique GO/EC terms” links, and click on the “Functional Annotations” in the left-hand menu.
7. To explore the species diversity, mouse-over the “Species Diversity” pie chart and click on the “Unique Species” link underneath for more details.

4.3 Using CATH to Perform Structure-Based Comparisons

CATHEDRAL can be used to perform a structural search of a query structure against CATH. Figure 11 describes the steps taken in an example scan with a query structure to search for structural matches to a CATH domain fold and homologs.

1. Click on the “Search” button on the CATH-Gene3D home page.
2. Select “Search By Structure”.
3. Upload a query PDB file and click “Submit” or “Reset” to start again. In this example, we will submit the PDB file of the recently deposited crystal structure, 4RGH (*see* Fig. 11a).
4. On the next page there is an option for users to select a particular chain to be scanned (if the protein has multiple chains) against CATH (*see* Fig. 11b). Select a chain to be scanned and/or view the match already found using a sequence

The figure is a composite of four panels (A, B, C, D) showing the CATH search workflow. Panel A shows the 'Search CATH' page with a file upload button and search options. Panel B shows 'Uploaded PDB Details' for PDB 4RGR, including title, header, chains, and a table of PDB chains with ATOM records, Residues, Sequence Hits, and Structural Scan results. Panel C shows the 'Scan submitted' confirmation page with progress indicators. Panel D shows 'PDB search results' for query structure 4RGR, featuring a scatter plot of SSAP score vs RMSD and a table of domain matches.

Table 1: PDB Chains (from Panel B)

Chain	ATOM records	Residues	Sequence Hits	Structural Scan
A	1001	128	Results available immediately	Scan took 4.56 seconds
B	943	120	Results available immediately	Scan took 4.56 seconds

Table 2: Domain Matches (from Panel D)

Domain	Length	Superfamily	RMSD	SSAP Score	Hits	Actions
4RGR_A	128	2.40.70.10	5.2	92.5	1	View details
4RGR_B	124	2.40.70.10	5.2	92.5	1	View details
4RGR_C	124	2.40.70.10	5.2	92.5	1	View details
4RGR_D	124	2.40.70.10	5.2	92.5	1	View details
4RGR_E	124	2.40.70.10	5.2	92.5	1	View details

Fig. 11 Performing a structural search against CATH, using as an example case PDB 4RGR, a recently deposited PDB structure (February 2016) downloaded from the PDB. (a) The PDB file is uploaded and submitted to the search tool. (b) The details of the uploaded PDB are provided, along with links to any sequence matches found in CATH. The protein is split into chains, and the user is given the option to run a structural scan against one or more of them. (c) Once the scan has been submitted, its progress can be monitored and/or the user can return to the previous page to submit a scan for another chain. (d) The CATHEDRAL scan results are shown for chain A and the most confident domain matches, ordered by decreasing SSAP score. Structural matches are colored *green* through to *red* to indicate their significance based on their SSAP score and RMSD value: a match with a SSAP score above 80 and an RMSD below 5 Å is likely to be a homolog

comparison method. In this example, we will submit the structure of chain A for scanning.

- When a structure has been submitted for scanning, links are provided to monitor the progress of the scan and to allow the user to return to their uploaded data (*see* Fig. 11c). The latter takes the user back to **step 3**, where another chain can be selected for scanning, for example.
- When the scan results are ready, a page similar to Fig. 11d will be generated. A scatter plot is generated, where each colored dot represents a matching CATH domain and the position of the dot is based upon two different measures of structural similarity: RMSD on the x-axis and SSAP score on the y-axis. The best matching domain to the submitted query will have the lowest RMSD and the highest SSAP score. Matches are colored from green for the most confident results through to red for the least confident results with the size of the colored dot representing the percentage overlap for the alignment between the query and match structures. The details of each match are listed in a table below the graph. For example, the

most confident match for the example scan, chain A from PDB 4RGH, is the domain 3s8iA00 (belonging to the superfamily 2.40.70.10). This is colored green due to the aligned pair having a low RMSD of 0.9 and a high SSAP score of 95.0. This dot is large compared to others in the scatter plot to highlight good overlap between the query and match aligned structures.

7. To find more information about domains related to a matched domain, click on the “Show related domains” button in any of the table rows. Each related domain ID is listed together with an image of itself, the number of residues it contains, and its full classification (i.e., both the CATH and SOLID levels). Clicking on the “Download” button downloads this information about all of the related domains. Clicking on a domain ID within the list opens the summary web page for that domain.

4.4 Sequence-Based Searches of CATH Superfamilies

Users can submit their own sequence of interest to CATH and run a sequence search to identify previously classified homologous domains and functional annotations based on the Functional Families (FunFams, *see* explanation in Subheading 2.4) to which constituent domains in the query sequence may belong (*see* Note 3).

1. Click on the “Search” button on the CATH-Gene3D homepage.
2. Select “Search By Sequence”. Enter a sequence in FASTA format or sequence identifier. Several examples are provided under “Examples.” The example “FASTA sequence” is shown in Fig. 12a. Click the green “Search” button (*see* Fig. 12a).
3. Following submission, the progress of the scan will be reported. A tick will appear at each of the four stages (shown in Fig. 12b) upon completion and the pipeline section will turn green. When the final stage of the scan has been completed, the CATH structural domain matches and the CATH functional family matches found can be viewed using the respective “Found X matches” buttons, where X equals the number of matches found (*see* Fig. 12b).
4. Clicking of the “Found 47 matches” button loads the functional family matches results page, which displays the most confident domain matches across the query sequence (if there are significantly scoring matches, *see* Fig. 12c). Each domain identified (that belongs to a functional family) is represented by a different row and may be discontinuous. Information provided for each domain match includes: the domain boundary positions, its superfamily and FunFam IDs, and the E-value.
5. To find out more about a domain match, click on a functional family name of interest within the “Match” column.

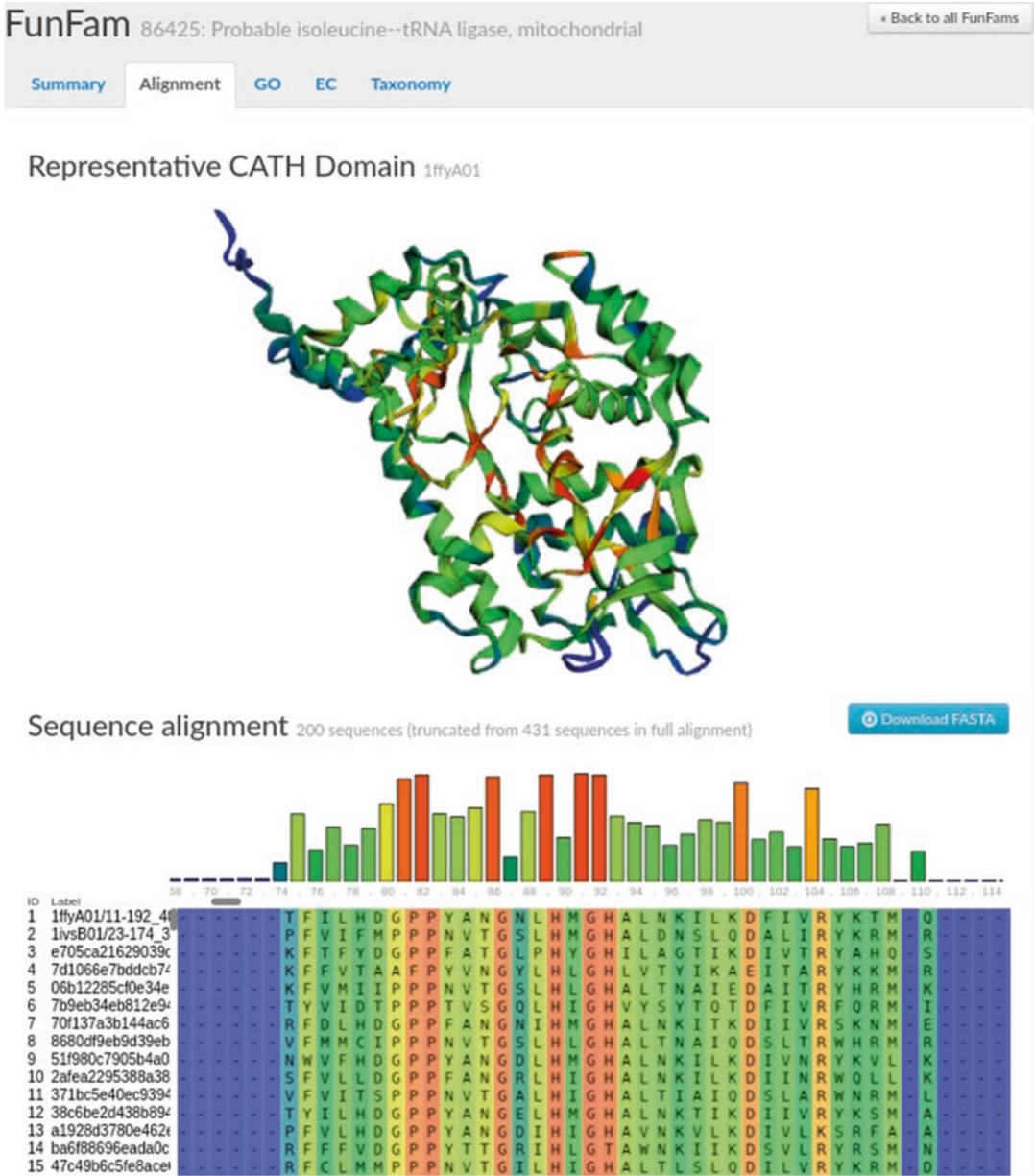


Fig. 13 The “Alignment” tab of a FunFam web page, which displays the interactive 3D structure of the representative CATH domain in the context of the whole PDB structure. Below the structure is the FunFam multiple sequence alignment, truncated to up to 200 sequences (for visualization purposes) where the highly conserved positions are colored *red*

sequence alignment are highlighted in red (i.e., alignment positions with a conservation score greater than 0.7, calculated with Scorecons).

- To download the FunFam alignment, click on the “Download FASTA” button.

5 Other Resources Using CATH-Gene3D Data

FunTree (www.funtree.info) is a resource developed by Furnham and Thornton [52] which exploits domain information from CATH superfamilies. FunTree allows users to explore the evolution of enzyme function through sequence, structure, phylogenetic, and functional information. Structural protein domains from CATH are searched against the MACiE database [53] to identify those with enzymatic function. MACiE uses expert manual curation to identify residues involved in catalytic functions. FunTree processes all the CATH-Gene3D domains containing these catalytic residues (e.g., both the domains with experimentally solved structure and predicted domains). Since some enzyme superfamilies in CATH contain very large and diverse set of sequences, the sequences are filtered by taxonomic lineage and unique function annotation before a phylogenetic tree is built, making the resulting tree easier to examine. Enzyme function is annotated using EC numbers. The similarity between pairs of EC numbers has been calculated using the EC-BLAST pipeline [54] to allow for the quantification of reaction mechanism similarity. The resource currently consists of 2340 CATH-Gene3D superfamilies, with over 70,000 structural domains and over 2300 EC numbers.

ArchSchema developed by Laskowski and Thornton [55] generates a two-dimensional network of multi-domain architectures (MDAs) identified for each domain representative in a given superfamily. Each node in the graph represents an MDA and the edges indicate those MDAs most closely related (i.e., sharing common domains), calculated using a simple Needleman and Wunsch-based alignment of domain constituents. The “parent” MDA is provided at the center of the graph with a gray background and domains are color-coded.

6 Concluding Remarks

The CATH-Gene3D resource, established in the mid-1990s, provides high-quality information on the classification of protein domains. A range of algorithms have been developed for classifying domains from protein structures deposited in the PDB into CATH superfamilies. Subsequently, sequence-based protocols are used to classify sequences from UniProt into these superfamilies (CATH-Gene3D). Thus, the classification pipeline involves a number of methods for comparing protein structures and for scanning sequences against HMMs for CATH superfamilies to validate homology (*see*, for example, the methods used to identify protein domains within a newly deposited protein chain from the Protein Data Bank, Subheading 3.1). Considerable manual curation is

necessary to classify remote homologs into CATH superfamilies. CATH-Gene3D classification data is publicly available and several search tools are available to allow users to research a protein(s) of interest. In particular, it is valuable in providing an overview of currently known folds in the protein structure universe, predicting which structural domains a protein sequence may contain, predicting the possible functions of a protein sequence, providing insights into distant evolutionary relationships, and providing highly curated gold-standard datasets, e.g., fold libraries for protein structure prediction methods.

7 Notes

1. While CATH covers a very reasonable amount of the protein structures in the PDB, it does not yet have complete coverage as extensive manual curation is required to validate very remote homologs. CATH contains nearly 80 % of all protein domains from completed genomes, where the remaining 20 % are mostly transmembrane-related and have very few structural arrangements [56]. However, coverage of the PDB has been significantly improved through the provision of “CATH-B,” which is a daily update of CATH providing putative domain boundary annotations and domain superfamily assignments. These data are available to download through: <http://www.cathdb.info/download>. While data in CATH-B may change in the next official release of CATH revisions are expected to be rare and as a result of additional evolutionary relationships detected (rather than any relationships removed). As the process of creating a new release of CATH involves numerous steps and therefore releases generally only occur annually, this daily update also allows users direct access to new data.
2. Some remote homologies between domains may be missed either through lack of evidence, or because the sequences or structures have diverged too far and the current technologies are not sufficiently sensitive. At every release of the CATH database, we attempt to identify missed homologies by performing a check called “Cross hits.” Representative domains from each superfamily are scanned against all other superfamilies to identify domain pairs that are potentially homologous (through sequence and structure similarity). The list of potentially homologous domains is then manually checked to confirm or reject the relationship.
3. Not all superfamily domain members belong to a FunFam. As superfamilies are subclassified using seed sequences that have at least one experimental functional annotation in the GO

resource [57], there may be functionally uncharacterized FunFams within the superfamily that are not currently classified. Due to the ever-increasing amount of sequence data it will always be extremely difficult to provide experimental annotations for all diverse functional relatives; however, more sequences will be annotated over time, which will allow us to subclassify and identify more FunFams. Functional subclassification of CATH superfamilies is performed with each new release to identify further FunFams that have acquired experimental characterization.

Acknowledgments

N.L.D. acknowledges funding from the Wellcome Trust (Award number: 104960/Z/14/Z). I.S. acknowledges funding from the BBSRC (Award number: BB/K020013/1). J.G.L. acknowledges funding from the BBSRC (Award number: BB/L002817/1). S.D.L. acknowledges funding from the Malaysian Ministry of Education.

References

- Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR et al (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112(3):535–542 <http://view.ncbi.nlm.nih.gov/pubmed/875032>
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. *Structure* 5(8):1093–1108 <http://www.ncbi.nlm.nih.gov/pubmed/9309224>
- Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(4):536–540 [citeulike-article-id:2564113](http://view.ncbi.nlm.nih.gov/pubmed/2564113)
- Oates ME, Stahlhacke J, Vavoulis DV, Smithers B, Rackham OJL, Sardar AJ et al (2015) The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Res* 43(D1):D227–D333 <http://dx.doi.org/10.1093/nar/gku1041>. Oxford University Press
- Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res* 42(D1):D310–D314 Oxford University Press <http://dx.doi.org/10.1093/nar/gkt1242>
- Fox NK, Brenner SE, Chandonia J-MM. 2014 SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 42(Database issue):D304–D309 Oxford University Press <http://dx.doi.org/10.1093/nar/gkt1240>
- Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S et al (2014) ECOD: an evolutionary classification of protein domains. *PLoS Comput Biol* 10(12):e1003926 <http://dx.doi.org/10.1371/journal.pcbi.1003926>. Public Library of Science
- Ekman D, Björklund ÅK, Frey-Skött J, Elofsson A (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* 348(1):231–243 <http://dx.doi.org/10.1016/j.jmb.2005.02.007>
- Holland TA, Veretnik S, Shindyalov IN, Bourne PE (2006) Partitioning protein structures into domains: why is it so difficult? *J Mol Biol* 361(3):562–590 <http://www.ncbi.nlm.nih.gov/pubmed/16863650>
- Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14(10):846–856 <http://www.ncbi.nlm.nih.gov/pubmed/9927713>
- Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M et al (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* 53(Suppl 6):491–496

- <http://www.ncbi.nlm.nih.gov/pubmed/14579338>
12. Taylor W, Orengo CA (1989) Protein structure alignment. *J Mol Biol* 208(1):1–22 [http://dx.doi.org/10.1016/0022-2836\(89\)90084-3](http://dx.doi.org/10.1016/0022-2836(89)90084-3)
 13. Orengo CA, Taylor WR (1996) [36] SSAP: Sequential structure alignment program for protein structure comparison. In: *Computer methods for macromolecular sequence analysis*. Elsevier, pp 617–635 [http://dx.doi.org/10.1016/S0076-6879\(96\)66038-8](http://dx.doi.org/10.1016/S0076-6879(96)66038-8)
 14. Swindells MB (1995) A procedure for detecting structural domains in proteins. *Protein Sci* 4(1):103–112 <http://dx.doi.org/10.1002/pro.5560040113>
 15. Siddiqui AS, Barton GJ (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci* 4(5):872–884 <http://dx.doi.org/10.1002/pro.5560040507>
 16. Holm L, Sander C (1994) Parser for protein folding units. *Proteins* 19(3):256–268 <http://dx.doi.org/10.1002/prot.340190309>
 17. Swindells MB (1995) A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci* 4(1):93–102 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2142969&tool=pmcentrez&rendertype=abstract>
 18. Rossmann MG, Liljas A (1974) Letter: recognition of structural domains in globular proteins. *J Mol Biol* 85(1):177–181 <http://www.ncbi.nlm.nih.gov/pubmed/4365123>
 19. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M et al (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35(Database issue):D291–D297 <http://dx.doi.org/10.1093/nar/gkl959>. Oxford University Press
 20. Orengo CA, Thornton JM (2005) Protein families and their evolution—a structural perspective. *Annu Rev Biochem* 74(1):867–900 <http://dx.doi.org/10.1146/annurev.biochem.74.082803.133029>. Department of Biochemistry and Molecular Biology, University College, London WC1E 6BT, United Kingdom. orengo@biochemistry.ucl.ac.uk
 21. Redfern OC, Harrison A, Dallman T, Pearl FMG, Orengo CA (2007) Cathedral: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol* 3:e232+ <http://dx.plos.org/10.1371%2Fjournal.pcbi.0030232>
 22. Subbiah S, Laurents DV, Levitt M (1993) Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr Biol* 3(3):141–148 [http://dx.doi.org/10.1016/0960-9822\(93\)90255-M](http://dx.doi.org/10.1016/0960-9822(93)90255-M)
 23. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33(7):2302–2309 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1084323&tool=pmcentrez&rendertype=abstract>
 24. Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol* 346(4):1173–1188 <http://dx.doi.org/10.1016/j.jmb.2004.12.032>. Department of Structural Biology, Fairchild Building, Stanford University, Stanford CA 94305, USA. trachel@cs.stanford.edu
 25. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951–960 <http://dx.doi.org/10.1093/bioinformatics/bti125>. Oxford University Press
 26. Eddy SR (1996) Hidden Markov models. *Curr Opin Struct Biol* 6(3):361–365 <http://www.ncbi.nlm.nih.gov/pubmed/8804822>
 27. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, et al 2015 HMMER web server: 2015 update. *Nucleic Acids Res* 43(W1):W30–W38. <http://nar.oxfordjournals.org/content/43/W1/W30>. Oxford University Press
 28. The UniProt Consortium. (2014). UniProt: a hub for protein information. *Nucleic Acids Res* 43(D1):D204–D212 <http://nar.oxfordjournals.org/content/43/D1/D204>
 29. Madera M (2008) Profilecomparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics* 24(22):2630–2631 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2579712\[&\]tool=pmcentrez\[&\]rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2579712[&]tool=pmcentrez[&]rendertype=abstract). Oxford Univ Press
 30. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33(Web Server issue):W244–W248 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1160169\[&\]tool=pmcentrez\[&\]rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1160169[&]tool=pmcentrez[&]rendertype=abstract)
 31. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL et al (2015) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44: D279–D285
 32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402 <http://dx.doi.org/10.1093/nar/25.17.3389>

- <http://dx.doi.org/10.1093/nar/25.17.3389>. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. altschul@ncbi.nlm.nih.gov:Oxford University Press
33. Lee DA, Rentzsch R, Orengo C (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Res* 38(3):720–737 <http://dx.doi.org/10.1093/nar/gkp1049>
 34. Capra JA, Singh M (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics* 24(13):1473–1480 <http://dx.doi.org/10.1093/bioinformatics/btn214> Oxford University Press
 35. Valdar WSJ (2002) Scoring residue conservation. *Proteins* 48(2):227–241 <http://dx.doi.org/10.1002/prot.10146>. Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, London, United Kingdom: Wiley Subscription Services, Inc., A Wiley Company
 36. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8(3):275–282 <http://view.ncbi.nlm.nih.gov/pubmed/1633570>. Department of Biochemistry and Molecular Biology, University College, London, UK
 37. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779–815 <http://www.ncbi.nlm.nih.gov/pubmed/8254673>
 38. Webb B, Sali A (2014) Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics* 47:5.6.1–5.6.32 <http://www.ncbi.nlm.nih.gov/pubmed/25199792>
 39. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, et al (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. <http://arxiv.org/abs/1601.00891>
 40. Moya Garcia A, Dawson NL, Kruger FA, et al (2016) A Structural and Functional View of Polypharmacology. *bioRxiv*
 41. Eddy SR (2011) Accelerated profile HMM searches. *PLoS Comput Biol* 7(10):e1002195 <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002195>
 42. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J et al (2013) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res* 41(Database issue):D483–D489 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531078&tool=pmcentrez&rendertype=abstract>
 43. Dessailly BH, Dawson NL, Mizuguchi K, Orengo CA (2013) Functional site plasticity in domain superfamilies. *Biochim Biophys Acta* 1834(5):874–889
 44. Yeats C, Redfern OC, Orengo C (2010) A fast and automated solution for accurately resolving protein domain architectures. *Bioinformatics* 26(6):745–751 <http://dx.doi.org/10.1093/bioinformatics/btq034>
 45. Lam SD, Dawson NL, Das S, Sillitoe I, Ashford P, Lee D et al (2016) Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res* 44(D1):D404–D409 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4702871&tool=pmcentrez&rendertype=abstract>
 46. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA (2015) Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics* 31(21):3460–3467 <http://bioinformatics.oxfordjournals.org/content/31/21/3460.abstract>. Oxford University Press
 47. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et al (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803857&tool=pmcentrez&rendertype=abstract>
 48. Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S et al (2014) Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic Acids Res* 42(Database issue):D240–D245 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965083&tool=pmcentrez&rendertype=abstract>
 49. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y et al (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 42(Database issue):D1091–D1097 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965102&tool=pmcentrez&rendertype=abstract>
 50. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C et al (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 40(Database issue):D841–D846 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245075&tool=pmcentrez&rendertype=abstract>
 51. Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6(7):e21800 <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021800>
 52. Furnham N, Sillitoe I, Holliday GL, Cuff AL, Rahman SA, Laskowski RA et al (2012) FunTree:

- a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res* 40(Database issue):D776–D782 <http://dx.doi.org/10.1093/nar/gkr852>Oxford University Press
53. Holliday GL, Almonacid DE, Bartlett GJ, O’Boyle NM, Torrance JW, Murray-Rust P et al (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res* 35(Database issue):D515–D520 [http://nar.oxfordjournals.org/content/35/suppl\[_\]1/D515.short](http://nar.oxfordjournals.org/content/35/suppl[_]1/D515.short)
 54. Rahman SA, Cuesta SM, Furnham N, Holliday GL, Thornton JM. (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat Methods* 11(2):171–174 <http://dx.doi.org/10.1038/nmeth.2803>. Nature Publishing Group.
 55. Tamuri AU, Laskowski RA (2010) ArchSchema: a tool for interactive graphing of related Pfam domain architectures. *Bioinformatics* 26(9):1260–1261 <http://www.ncbi.nlm.nih.gov/pubmed/20299327>
 56. Sillitoe I, Dawson N, Thornton J, Orengo C (2015) The history of the CATH structural classification of protein domains. *Biochimie* <http://www.sciencedirect.com/science/article/pii/S0300908415002515>
 57. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* 25(1):25–29 Nature Publishing Group

Structure-Based Virtual Screening

Qingliang Li and Salim Shah

Abstract

Structure-based virtual screening (SBVS) is a computational approach used in the early-stage drug discovery campaign to search a chemical compound library for novel bioactive molecules against a certain drug target. It utilizes the three-dimensional (3D) structure of the biological target, obtained from X-ray, NMR, or computational modeling, to dock a collection of chemical compounds into the binding site and select a subset of these compounds based on the predicted binding scores for further biological evaluation. In the present work, we illustrate the basic process of conducting a SBVS with examples using freely accessible tools and resources.

Key words Molecular docking, Virtual screening, Drug discovery, UCSF Chimera, AutoDock Vina, OpenBabel

1 Introduction

Virtual screening (VS) is a computational way to search a chemical compound database to identify novel molecules with required biological activity. It is usually described as a cascade of sequential filters to narrow down a large number of compounds to a small set of hits with potential bioactivity against a certain drug target. The term “virtual screening” was coined in the late 1990s [1, 2]. Many tools and techniques have been developed since then [3–5]. VS has become a valuable approach as a complement to high-throughput screening (HTS) in the pharmaceutical industry, as well as in small biotechnology companies and academic labs.

In general, VS can be classified into two categories, ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS) [2, 6, 7]. LBVS uses the structure-activity data from a set of known active compounds to identify candidate compounds for experimental evaluation, including structural similarity searching, quantitative structure-activity relationship (QSAR) studies, pharmacophore, or 3D shape matching [8, 9]. On the other hand, SBVS utilizes the 3D structure of the biological target, obtained

from X-ray, NMR, or computational modeling, to dock candidate compounds into the binding site and then rank them based on the predicted binding affinity and/or complementarity scores.

SBVS comprises three pillars: (1) molecular docking, (2) receptor 3D structure, and (3) chemical compound library.

Molecular docking, the core technique of SBVS, intends to predict the structure of the intermolecular complex formed between two or more constituent molecules, typically a small molecule and a protein target [10]. Docking protocols can be described as a combination of two complementary components: a search algorithm and a scoring function. The search algorithm, such as Genetic Algorithm or Monte Carlo search, entails exploring the conformational space of the ligand and/or the receptor (sampling) [11]; while the scoring function, including force-field-based, empirical and knowledge-based functions, entails evaluating each of the “sampled” binding modes (scoring) [12]. Since the pioneering work of molecular docking in the early 1980s [13], many docking tools have been developed [10, 11], including DOCK [14, 15], AutoDock [16], AutoDock Vina [17], Glide [18], and GOLD [19, 20]. Despite the invention of these state-of-the-art techniques, the great challenge for docking is that no single tool can consistently predict correct bioactive compounds from a large chemical library, and this is also the reason why the true potential of VS is questioned [1, 3, 21].

Receptor 3D structures are usually retrieved from the Protein Data Bank (PDB), which is a structural database of large biological molecules, such as proteins and nucleic acids, determined by X-ray crystallography or NMR spectroscopy [22]. Advances in structural biology and structural genomics in the past decade have generated a great many 3D structures of biologically relevant macromolecules, including potential therapeutic targets, offering unparalleled opportunities for structure-based virtual screening and drug discovery. At the time of this work, there are more than 114,000 biological macromolecular structures available in the PDB and it continues to increase every year.

A chemical compound library is a collection of chemical compounds used for SBVS. Currently, a lot of public and commercial chemical databases are available, such as PubChem [23, 24], ChEMBL [25], ZINC [26], ChemSpider [27], ChemBridge (<http://chembridge.com>) and Maybridge (<http://www.maybridge.com>). These resources dramatically lower the barrier to entry into drug development for researchers in academia and small biotech companies.

A typical workflow of SBVS usually starts with the preparation of the receptor structure and chemical library, which are merged at the molecular docking step (Fig. 1). From docking, each compound will obtain a docking score, indicating its potential to be active against the receptor. By ranking all the compounds based on

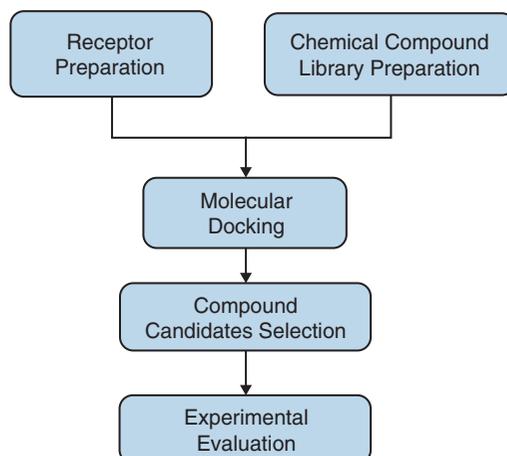


Fig. 1 Typical workflow for structure-based virtual screening (SBVS)

their docking scores, a small set of compounds with better scores will survive for further experimental evaluation. The workflow may have different variants for specific requirements by incorporating additional filters to increase the success rate, such as adding consensus scoring after docking, adding a drug-likeness filter in chemical library preparation, etc.

In the present work, we focus on how to set up and run the basic SBVS workflow with simple examples. All software and tools used here are freely accessible, so readers may have a try while reading. In Subheading 3, we start with one molecule docking with minimum software installation requirements—only UCSF chimera. To dock multiple compounds, we assume that readers have knowledge of the basic Unix/Linux command line environment. In Subheading 4 (Notes), we will discuss potential problems that readers may have in practice.

2 Materials

1. Chemical compound structures obtained from PubChem [23, 24] (<https://pubchem.ncbi.nlm.nih.gov>), which is a public database of chemical molecules and their bioactivities.
2. Receptor 3D structures obtained from the PDB (<http://www.rcsb.org>), which is a public database for 3D structural data of large biological molecules, such as proteins, DNA, and RNA.
3. Software and tools:
 - (a) UCSF Chimera [28] (<https://www.cgl.ucsf.edu/chimera>), which is a visualization tool for molecular modeling and structural biology.
 - (b) AutoDock Vina [17] (<http://vina.scripps.edu>), which is an open-source program for molecular docking.

- (c) OpenBabel [29] (<http://openbabel.org>), which is free software mainly used for converting chemical file formats.

UCSF Chimera has a graphic interface for AutoDock Vina, in which there is an online version of AutoDock Vina provided by Opal web service (<http://nbc.ucsdc.edu/data/docs/opal/index.html>). As it currently only supports docking one compound at a time within Chimera, you may use it to finish the Quickstart in Subheading 3 without installing AutoDock Vina locally. However, if you want to dock multiple compounds or carry out real VS, you are required to be familiar with basic Unix/Linux commands and able to install the software and tools. For Mac and Unix/Linux users, all commands can be run in the preinstalled terminal application of the operating system. For windows users, Cygwin (<https://www.cygwin.com>) can be used to simulate the Unix/Linux environment.

3 Methods

Here, we use nuclear receptor-binding SET domain protein 1 (NSD1) as an example to screen a collection of chemical compounds. NSD1 is a histone methyltransferase that regulates the activity of genes involved growth and development [30]. At the time of this work, one crystal structure of NSD1 (PDB: 3ooi) in complex with its native ligand, S-Adenosylmethionine (SAM), was available in the PDB, and a known inhibitor of NSD1, Sinefungin (SIN), was available in PubChem (CID: 65482).

3.1 Quickstart— Dock One Chemical Compound Against the Receptor Using UCSF Chimera

3.1.1 Receptor Structure Preparation

1. Obtain the receptor structure. (In Chimera: File → Fetch by ID... → select PDB and fill in ID with 3ooi) (Fig. 2).
2. Clean the receptor structure. Open the “Dock Prep” window (In Chimera: Tools → Surface/Binding Analysis → Dock Prep) and select “3ooi” in “Molecules to prep” and check all the checkboxes except the last two: “Add charges” and “Write Mol2 file” (Fig. 3). Click “OK” to open the “Add Hydrogens for Dock Prep” window. In the window, keep the default setting unchanged and click “OK” (Fig. 4). Additionally, delete the two sulfates in the structure: (1) select the sulfates (in Chimera: Select → Residues → SO4); (2) delete them (in Chimera: Actions → Atoms/Bonds → delete) (*see Note 1* for other issues related to structure preparation).
3. Split the receptor and ligand if applicable. Open the command line (in Chimera: Tools → General Controls → Command line) and enter the following command between the quotes: “split #0 ligands.” In the Model Panel (in Chimera: Tools → General Controls → Model Panel), the original structure (#0) will be split into two models, “#0.1 3ooi” and “#0.2 3ooi SAM,” representing the NSD1 receptor and SAM ligand, respectively.

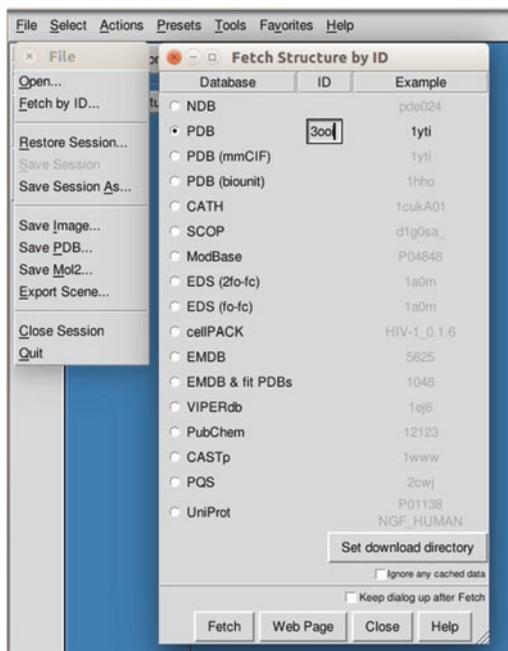


Fig. 2 Retrieve receptor 3D structure from the PDB in UCSF Chimera

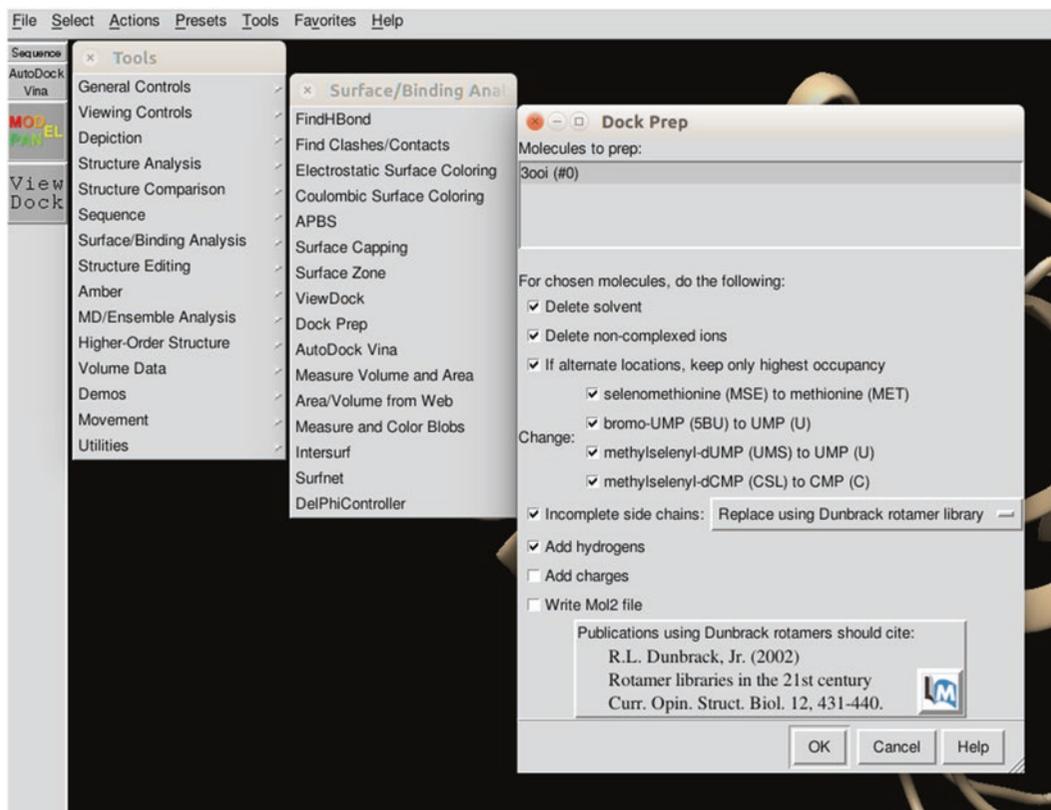


Fig. 3 Clean and repair receptor 3D structure in UCSF Chimera

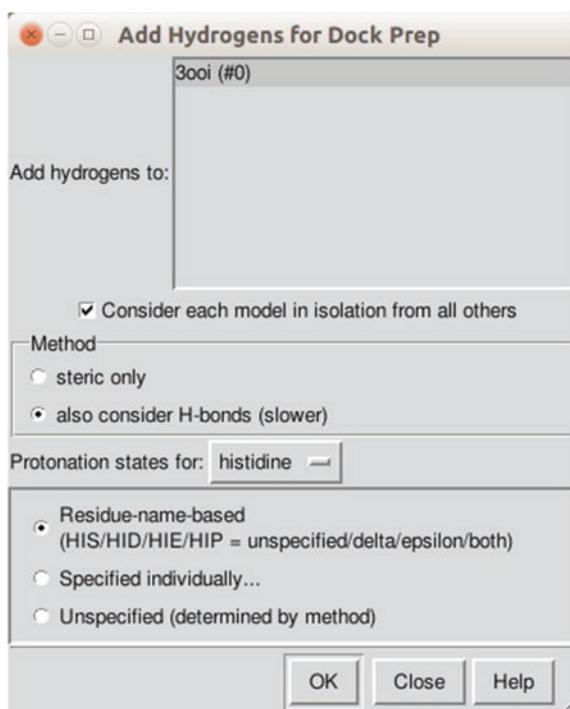


Fig. 4 Add hydrogens to receptor 3D structure in UCSF Chimera

3.1.2 Chemical Compound Preparation

1. Obtain the chemical compound structure. Open “Fetch by ID...” (in Chimera: File → Fetch by ID...), select PubChem and fill in ID with 65482. At the time of this work, there was a bug in fetching PubChem compounds. In this case, you can always use an alternative way to directly download the structure and open it in Chimera. Use the following link to download the 3D structure of SIN (PubChem ID: 65482): http://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/65482/SDF?record_type=3d. When opening the downloaded structure in Chimera, you will see a new model (ID: 1, name 65482) in the Model Panel. For clarity, you may rename it SIN by using the “rename” button listed on the right side.

3.1.3 Dock the Chemical Compound Against the Receptor

1. Open the “AutoDock Vina” interface (In Chimera: Tools → Surface/Binding Analysis → AutoDock Vina, Fig. 5).
2. In the “Output file” field, select a folder (by clicking the “Browse” button) to save all the docking files. Then, in the popup window (“Save File in Chimera”), input “eg1” in the “File Name” field.
3. In the “Receptor” field, select “3ooi (#0.1)” from the dropdown list.
4. In the “Ligand” field, select “SIN (#1)” from the dropdown list.

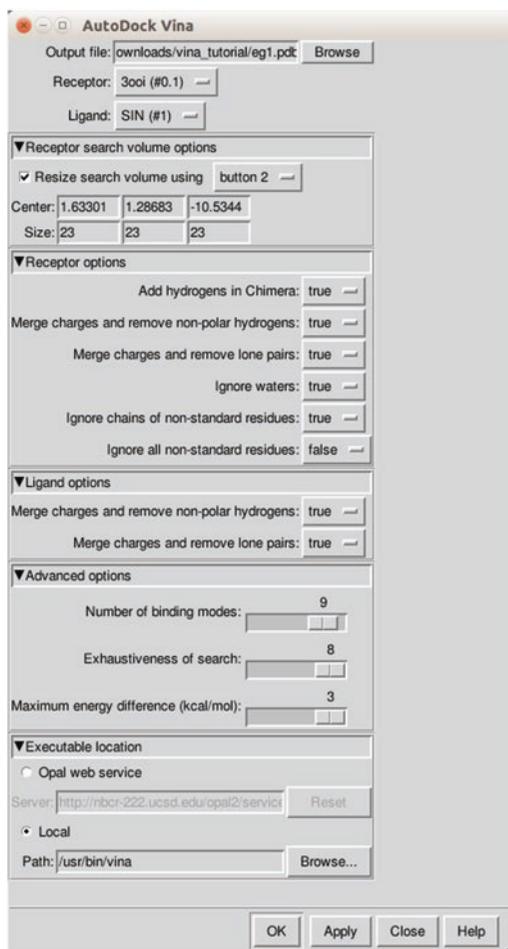


Fig. 5 Autodock Vina settings in UCSF Chimera

5. In the “Receptor search volume options” section, check the box “Resize search volume using button 2” (middle button of your mouse). Then click any visualization area in the main window in Chimera with your “button 2,” and a green box will appear. You may change the size and location of the box to define the searching area for docking. The box is required to enclose the binding site area, i.e., around SAM. You may see the coordinates of “Center” and “Size” change accordingly along with the box. You can also use the coordinates (Center: 1.63301, 1.28683, -10.5344; Size: 23, 23, 23) by manually inputting them in the fields for this exercise (For issues related to how to set the searching box, *see* **Note 2**).
6. In the “Receptor options” section, keep default settings (*see* **Note 3**).
7. In the “Ligand options” section, keep default settings (*see* **Note 3**).

8. In the “Advanced options” section, keep default settings (*see* **Note 4**).
9. In the “Executable location” section, you may use the online version of AutoDock Vina (provided by Opal web service) or your local installation (specifying the location of AutoDock Vina).
10. Click “Apply” to run the docking. The docking status will be shown on status bar in the bottom of Chimera main window. When it finishes, a “ViewDock” window will show all the docking modes and scores.
11. A successful docking will generate six files in the “Output file” folder:
 - (a) “eg1.pdbqt” contains the nine best binding modes of SIN.
 - (b) “eg1.conf,” contains the docking parameters for AutoDock Vina, such as searching box size and location, number of output binding modes etc.
 - (c) “eg1.receptor.pdb” is the receptor structure in PDB format to be converted into a PDBQT file.
 - (d) “eg1.receptor.pdbqt” is the prepared receptor structure in PDBQT format that can be directly used by AutoDock Vina.
 - (e) “eg1.ligand.pdb” is the compound structure in PDB format to be converted into a PDBQT file.
 - (f) “eg1.ligand.pdbqt” is the prepared compound structure in PDBQT format that can be directly used by AutoDock Vina.

3.1.4 Visualize the Docking Results

1. The ViewDock window will automatically be open when docking is completed (Fig. 6). You may also manually start it (in Chimera: Tools → Surface/Binding Analysis → ViewDock) and select the docking results (“eg1.pdbqt”) in the “Output file” folder you specified in **step 2** of Subheading 3.1.3.
2. You may browse individual binding modes of SIN in ViewDock.
 - (a) The “Score” column shows the predicted binding affinity (in kcal/mol), namely the docking score, which is used to rank chemical compounds in SBVS.
 - (b) The columns “RMSD ub” (upper bound) and “RMSD lb” (lower bound) are two variants of the root-mean-square deviation (RMSD) values that measure the match relative to the best binding mode using only movable heavy atoms. According to AutoDock Vina’s manual:
 - RMSD/ub matches each atom in one conformation with itself in the other conformation, ignoring any symmetry.



Fig. 6 Visualizing Autodock Vina docking results in UCSF Chimera

- RMSD' matches each atom in one conformation with the closest atom of the same element type in the other conformation (RMSD' cannot be used directly, because it is not symmetric).
- RMSD/lb is defined as follows: $\text{RMSD}/\text{lb}(c_1, c_2) = \max(\text{RMSD}'(c_1, c_2), \text{RMSD}'(c_2, c_1))$.

3.2 Virtual Screening—Dock Multiple Chemicals Using the Command Line Environment

3.2.1 Receptor Structure Preparation

If you have completed Subheading 3.1.3 of the “Quickstart” example, you can find a file named “eg1.receptor.pdbqt” in the “Output file” folder. The file is the prepared receptor file ready for AutoDock Vina docking. Thus, you can skip this section and go to Subheading 3.2.2 directly.

If you have not completed Subheading 3.1.3 of the “Quickstart” example, you may save the “3ooi (#0.1)” model obtained at the end of Subheading 3.1.1 into a file named “eg1.receptor.pdb.” Then, use the script `pdb2pdbqt.sh` (<https://github.com/leonqli/useful-scripts-for-autodock-vina/blob/master/pdb2pdbqt.sh>) to prepare the receptor file (eg1.receptor.pdbqt) as follows:

```
$chmod +x pdb2pdbqt.sh
$bash pdb2pdbqt.sh eg1.receptor.pdb eg1.receptor.pdbqt
```

3.2.2 Chemical Compounds Preparation

At the beginning of this section, we assume all chemical compounds used here are: (1) in 3D, namely that all atoms have reasonable x , y , z coordinates; (2) in SDF format (https://en.wikipedia.org/wiki/Chemical_table_file) (see **Note 5** for how to

generate 3D molecules and **Note 6** for chemical library preparation).

1. Download ten compounds from DTP/NCI in PubChem using this link (http://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/90945205,85090241,60148220,60148180,60147857,60147849,60147847,57008769,54723763,54723759/SDF?record_type=3d) and save it with the name, “ligs.sdf.” Please make sure to specify 3D (record_type=3d) when downloading the SDF file.
2. As the ligand preparation script, prepare_ligand4.py, accepts one compound per file in either Mol2 or PDB format as input, use OpenBabel to convert chemical compounds from SDF to Mol2 and split them into individual file with the option “-m.” At the same time, add hydrogens to each compound, if applicable, with the option “-h”:

```
$mkdir ligs_mol2
$babel -isdf ligs.sdf -omol2 ligs_mol2/lig.
mol2 -h -m
```

At this step, ten Mol2 files, from lig1.mol2 to lig10.mol2, are generated in the ligs_mol2 folder.

3. Convert all chemical compounds in Mol2 format to PDBQT format in a batch mode with the script mol2s_to_pdbqts.sh (https://github.com/leonqli/useful-scripts-for-autodock-vina/blob/master/mol2s_to_pdbqts.sh):

```
$mkdir ligs_pdbqt
$chmod +x ./mol2s_to_pdbqts.sh
$./mol2s_to_pdbqts.sh ligs_mol2 ligs_pdbqt
```

At this step, a total of ten PDBQT files, from lig1.pdbqt to lig10.pdbqt, are generated in the ligs_pdbqt folder.

3.2.3 Dock Multiple Chemical Compounds Against the Receptor

Three files are required to run a docking with AutoDock Vina: (1) prepared receptor file; (2) prepared ligand file; and (3) AutoDock dock configuration file. Since you have prepared the first two in the previous steps, the only file needed is the configuration file that contains docking parameters.

1. Create the configuration file. If you run the docking in Quickstart, you may use the generated file, “eg1.conf” directly. Alternatively, you may manually create the file, containing the following content:

```
center_x = 1.63
center_y = 1.29
center_z = -10.53
```

```
size_x = 23.00
size_y = 23.00
size_z = 23.00

energy_range = 3
exhaustiveness = 8
num_modes = 9
```

To decide on the size and location of the searching box, please refer Subheading 3.1.3 in the Quickstart.

2. Dock the chemical compounds against the receptor using the script `dock.sh` (<https://github.com/leonqli/useful-scripts-for-autodock-vina/blob/master/dock.sh>):

```
$chmod +x dock.sh
$mkdir out
$mkdir log
$chmod +x dock.sh
$/dock.sh eg1.conf eg1.receptor.pdbqt ligs_
  pdbqt out log
```

At this step, the docking modes and logs of the compounds are generated in the folders “out” and “log,” respectively.

3. To view and rank all the docking scores, use the script `logs2csv.sh` (<https://github.com/leonqli/useful-scripts-for-autodock-vina/blob/master/logs2csv.sh>):

```
$chmod +x logs2csv.sh
$logs2csv.sh log scores.csv
```

The script will generate a csv file with ligand name and docking scores:

```
lig2, -8.8
lig8, -8.5
lig10, -8.4
lig1, -8.4
lig3, -8.1
lig4, -7.8
lig5, -7.6
lig7, -7.3
lig6, -6.3
lig9, -6.0
```

4. Select compound candidates based on the docking scores for further experimental evaluation. In most real SBVS campaigns, additional scoring and filtering approaches may be added at this step to increase the chances of success“ (see **Note 7**). Usually, manual inspection of the docking modes is required to finalize the compound candidates for experiments (see Subheading 3.1.4 for how to visualize docking results).

4 Notes

1. There are several issues when selecting and preparing the receptor structure: (1) receptor 3D structures from PDB or other resources may have different qualities or accuracies [31], thus always select high-quality structures [32] if possible, especially for the binding site region; (2) some water molecules or ions may coordinate correct ligand binding or mediate ligand-receptor interactions by forming hydrogen bonds; thus they should not be deleted, but kept as part of the receptor structure [21].
2. The searching box setting is critical for a successful docking. If the binding site is known (binding of an inhibitor, for example), set the box to enclose the entire inhibitor and add more space in all dimensions to ensure the free rotation of the ligand in it. To consider an extreme case, setting a big box enclosing the entire receptor will take a lot of computational time to dock. Moreover, if the docking is for VS, it may make it hard to interpret the docking results, because compounds may be docked to an unknown site on the receptor surface, which has yet to be verified as a true binding site. Thus, a tradeoff needs to be made when setting the searching box.
3. AutoDock Vina does not use charges or nonpolar hydrogens in the scoring function, so the settings are not expected to affect docking results. The AutoDock Vina interface in UCSF Chimera uses the AutoDock receptor and ligand preparation scripts, known as `prepare_receptor4.py` and `prepare_ligand4.py`, on the backend.
4. The maximum settings of the “Advanced options” are fixed for AutoDock Vina in the UCSF Chimera interface. These settings can be changed in the command line version; for example, the “Exhaustiveness of search” can be increased to more than eight.
5. AutoDock Vina requires 3D chemical compounds as input, namely each atom with x , y , z coordinates. There are several tools that can be used to generate 3D structures, including CORINA (<https://www.molecular-networks.com/products/corina>; commercial software), OpenBabel (relatively slow, not suitable for a large number of compounds), Balloon [33] and Frog 2 (webserver) [34].
6. Because large-scale docking is a time-consuming process and requires lots of computational resources, it usually uses certain filters to preselect the chemical library, such as drug-likeness filters (Lipinski’s rule of 5) [35]; structural similarity searching of known bioactive ligands to enrich the library; or selecting only purchasable compounds. Moreover, both SDF and

SMILES formats are usually used to store large-scale chemical compounds. SDF can store 2D and 3D chemical structures, while SMILES format can only store 0D structure (no coordinates). OpenBabel is a useful tool for converting file formats.

7. Because of the current limitation of docking methods, the prediction accuracy is still relatively low. In order to increase the success rate, many variants have been proposed with additional filters added into the workflow, such as consensus scoring with multiple scoring functions [36, 37], or combining ligand-based and structure-based VS [38]. In addition, incorporating receptor flexibility and/or ensemble docking, and/or selecting potential hits with diverse scaffolds for further experiments, are also used in SBVS campaigns [39–41].

References

1. Schneider G (2010) Virtual screening: an endless staircase? *Nat Rev Drug Discov* 9:273–276
2. Lavecchia A, Di Giovanni C (2013) Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem* 20:2839–2860
3. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* 11:580–594
4. Köppen H (2009) Virtual screening—what does it give us? *Curr Opin Drug Discov Devel* 12:397–407
5. Song CM, Lim SJ, Tong JC (2009) Recent advances in computer-aided drug design. *Brief Bioinform* 10:579–591
6. Heikamp K, Bajorath J (2013) The future of virtual compound screening. *Chem Biol Drug Des* 81:33–40
7. Muegge I, Oloff S (2006) Advances in virtual screening. *Drug Discov Today Technol* 3:405–411
8. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J Chem Inf Model* 50:205–216
9. Stahura FL, Bajorath J (2005) New methodologies for ligand-based virtual screening. *Curr Pharm Des* 11:1189–1202
10. Sousa SF, Fernandes PA, Ramos MJ (2006) Protein-ligand docking: current status and future challenges. *Proteins* 65:15–26
11. Moitessier N, Englebienne P, Lee D et al (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol* 153(Suppl 1):S7–S26
12. Huang S-Y, Grinter SZ, Zou X (2010) Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys Chem Chem Phys* 12:12899–12908
13. Kuntz ID, Blaney JM, Oatley SJ et al (1982) A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 161:269–288
14. Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* 15:411–428
15. Moustakas DT, Lang PT, Pegg S et al (2006) Development and validation of a modular, extensible docking program: DOCK 5. *J Comput Aided Mol Des* 20:601–619
16. Morris GM, Goodsell DS, Huey R, Olson AJ (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J Comput Aided Mol Des* 10:293–304
17. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31:455–461
18. Friesner RA, Banks JL, Murphy RB et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* 47:1739–1749
19. Jones G, Willett P, Glen RC et al (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 267:727–748
20. Jones G, Willett P, Glen RC (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol* 245:43–53

21. Cheng T, Li Q, Zhou Z et al (2012) Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J* 14:133–141
22. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
23. Li Q, Cheng T, Wang Y, Bryant SH (2010) PubChem as a public resource for drug discovery. *Drug Discov Today* 15(23–24):1052–1057. doi: [10.1016/j.drudis.2010.10.003](https://doi.org/10.1016/j.drudis.2010.10.003)
24. Wang Y, Xiao J, Suzek TO et al (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 37:W623–W633
25. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40:D1100–D1107
26. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45:177–182
27. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87:1123–1124
28. Pettersen EF, Goddard TD, Huang CC et al (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612
29. O’Boyle NM, Banck M, James CA et al (2011) Open Babel: an open chemical toolbox. *J Cheminform* 3:33
30. Wang GG, Cai L, Pasillas MP, Kamps MP (2007) NUP98–NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis. *Nat Cell Biol* 9:804–812
31. Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science* 294:93–96
32. Warren GL, Do TD, Kelley BP et al (2012) Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov Today* 17:1270–1281
33. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model* 47:2462–2474
34. Miteva MA, Guyon F, Tufféry P (2010) Frog2: efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res* 38:W622–W627
35. Lipinski CA (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* 44:235–249
36. Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* 42:5100–5109
37. Wang R, Wang S (2001) How does consensus scoring work for virtual library screening? An idealized computer experiment. *J Chem Inf Comput Sci* 41:1422–1426
38. Swann SL, Brown SP, Muchmore SW et al (2011) A unified, probabilistic framework for structure- and ligand-based virtual screening. *J Med Chem* 54:1223–1232
39. Osguthorpe DJ, Sherman W, Hagler AT (2012) Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. *J Phys Chem B* 116:6952–6959
40. Durrant JD, McCammon JA (2010) Computer-aided drug-discovery techniques that account for receptor flexibility. *Curr Opin Pharmacol* 10:770–774
41. Wale N, Karypis G, Watson IA (2007) Method for effective virtual screening and scaffold-hopping in chemical compounds. *Comput Syst Bioinformatics Conf* 6:403–414

Part II

Protein PTM (Post-Translational Modification) Bioinformatics

Bioinformatics Analysis of Protein Phosphorylation in Plant Systems Biology Using P3DB

Qiuming Yao and Dong Xu

Abstract

Protein phosphorylation is one of the most pervasive protein post-translational modification events in plant cells. It is involved in many plant biological processes, such as plant growth, organ development, and plant immunology, by regulating or switching signaling and metabolic pathways. High-throughput experimental methods like mass spectrometry can easily characterize hundreds to thousands of phosphorylation events in a single experiment. With the increasing volume of the data sets, Plant Protein Phosphorylation DataBase (P3DB, <http://p3db.org>) provides a comprehensive, systematic, and interactive online platform to deposit, query, analyze, and visualize these phosphorylation events in many plant species. It stores the protein phosphorylation sites in the context of identified mass spectra, phosphopeptides, and phosphoproteins contributed from various plant proteome studies. In addition, P3DB associates these plant phosphorylation sites to protein physicochemical information in the protein charts and tertiary structures, while various protein annotations from hierarchical kinase phosphatase families, protein domains, and gene ontology are also added into the database. P3DB not only provides rich information, but also interconnects and provides visualization of the data in networks, in systems biology context. Currently, P3DB includes the KiC (Kinase Client) assay network, the protein-protein interaction network, the kinase-substrate network, the phosphatase-substrate network, and the protein domain co-occurrence network. All of these are available to query for and visualize existing phosphorylation events. Although P3DB only hosts experimentally identified phosphorylation data, it provides a plant phosphorylation prediction model for any unknown queries on the fly. P3DB is an entry point to the plant phosphorylation community to deposit and visualize any customized data sets within this systems biology framework. Nowadays, P3DB has become one of the major bioinformatics platforms of protein phosphorylation in plant biology.

Key words Bioinformatics database, Domain co-occurrence network, Kinase-substrate network, Phosphatase-substrate network, Plant protein phosphorylation, Protein-protein interaction network, Systems biology

1 Introduction

Among the many kinds of protein post-translational modifications, protein phosphorylation is one of the most prevalent types in plant cells to control and regulate signaling and metabolic pathways. It plays critical roles in various plant biological processes and functionalities, including growth and development, functional traits,

nutrient metabolism [1], and immunity [2]. Therefore, protein phosphorylation is the most intensively studied protein modification type because of its ability to reveal molecular insights for signaling and gene regulation. With the constantly reducing cost of proteomic studies, experimental characterization of phosphorylation events has shifted from gel-based western blot to mass spectrometry (MS). With more and more genomes completely sequenced in plant species, proteome-wide phosphorylation identification is becoming a common practice. Currently, a single MS run with a high-performance spectrum search algorithm can identify hundreds to thousands of unique phosphorylated peptides of high confidence within several hours. With the rapidly increasing volume of phosphoproteome data sets, several phosphorylation databases have been created in the past decade, e.g., PhosphAT [3] and P3DB [4], specifically for the plant phosphoproteomes. PhosphAT is a comprehensive phosphorylation database for *Arabidopsis*, including a predictor and kinase-target information. P3DB contains model organisms and crop plants, i.e., *Arabidopsis thaliana*, *Brassica napus*, *Glycine max*, *Medicago truncatula*, *Nicotiana tabacum*, *Oryza sativa*, *Solanum tuberosum*, *Vitis vinifera*, and *Zea mays*. P3DB is a comprehensive, systematic, and interactive online platform to deposit, query, analyze, and visualize the phosphorylation events in plant species.

In order to organize and display the phosphorylation data at the systems level, P3DB adopts three layers of data integration. Each layer of data representation is accompanied with different functional or tool modules. We will briefly describe the basic design and the challenges of data integrations.

The first layer is the core phosphorylation data composed of peptide spectra, phosphopeptide, phosphosite, and phosphoproteins. In a proteomic data set, phosphoproteins usually contain multiple phosphosites, which are identified from one or more phosphopeptides in multiple spectra. P3DB designs hierarchical data pages to display proteomics consistent with both top-down and bottom-up approaches. The redundant data at the spectra level, phosphosite level, and phosphopeptide level can then be completely addressed in this hierarchical structure. Moreover, P3DB provides cross-references on the phosphoprotein or phosphopeptide page to integrate multiple results from different studies or publications. P3DB allows users to browse, search, or perform BLAST sequence similarity search on the core phosphorylation data. At this moment, P3DB is one of the very few databases incorporating detailed bottom-up proteomic data, especially the MS spectra for phosphorylation.

For data searches, P3DB generally provides the second layer of data, i.e., comprehensive protein annotations for phosphorylation events to serve as a one-stop data shop. The core phosphorylation data, especially phosphoproteins, are associated with multiple annotation elements, such as a position-based protein chart

showing physicochemical properties, an interacting protein list, gene ontology terms (GO terms), protein tertiary structures, orthologous sequences, and kinase or phosphatase classification, all of which can be essential to understanding the context for the phosphorylation event.

While all of the above information is displayed in a protein-centric data structure, P3DB also weaves proteins into networks. The KiC-assay network is created directly from in vitro kinase substrate screening by using synthetic peptides [5, 6]. Besides, P3DB integrates several classic networks into the platform, i.e., a protein-protein interaction (PPI) network, a protein domain co-occurrence network, a protein kinase-substrate network, and a protein phosphatase-substrate network. These networks facilitate the identification of potential interactions or complex partners, and the related functional domain topology helping the biologists to deduce or interpret the potential interactions or pathways in plants due to the phosphorylation event.

P3DB is, therefore, an ideal platform to query, browse, and visualize the plant phosphorylation data in multiple dimensions, especially in the systems biology context. P3DB is a reliable database that has been built by gathering and filtering information with high-quality data. Furthermore, P3DB is also designed for community-based data depository. Users can share their own data sets and automate data depository processes for publication or any peer preview purposes. This provides an alternative way for P3DB to collect data in the long run.

In this Chapter, we will present a major collection of functionalities in P3DB (Table 1) to illustrate the bioinformatics analyses

Table 1
Summary of toolkits in P3DB 3.5 (data type and associated tool for query and display)

Data	Tool
Core phosphorylation data	Search engine; BLAST; hierarchical display pages
Enzyme (kinase/phosphatase)	Search engine; family browser; kinase/phosphatase network
Gene ontology	Search engine; ontology browser
KiC assay data	KiC assay data display; network
Taxonomy	Taxonomy browser
Prediction data	Plant-specific phosphorylation predictor
Protein domain	Search engine; protein chart; domain network
Protein interaction	Search engine; PPI network
Protein physicochemical properties	Protein chart

that can be performed on protein phosphorylation data, from querying a single phosphorylation site to browsing networks. Materials, including data resources and toolkits used in P3DB, will be listed in the next Subheading. The detailed functionalities and services in P3DB will then be illustrated in the Subheading 3, and the Subheading 4 will introduce specific ideas or technical tricks.

2 Materials

P3DB website can be accessed at <http://p3db.org>. The data sets in the P3DB version 3.5 were curated from 32 publications covering nine plant species. P3DB version 3.5 now has 47,923 nonredundant phosphorylation sites in 16,477 phosphoproteins. The phosphoprotein distribution by species (in percentage) is 30.15 % in *Arabidopsis thaliana*, 25.36 % in *Medicago truncatula*, 29.31 % in *Oryza sativa*, and 15.18 % for other species. P3DB supports sharing private data sets in a controlled fashion, which is useful to preview the data sets within or among research groups in a set of customized data pages. These data sets can be merged into the public repository after the official review process has been completed and with the user's agreement.

P3DB mainly includes in vivo experimental data, which reflects the actual phosphorylation processes in the cell, together with some in vitro data (e.g., those obtained for kinase-client relationships, i.e., KiC-Assay data [6]). In addition, P3DB provides a plant-specific phosphorylation site predictor only for on-the-fly prediction.

The analysis of the raw MS spectra usually differs across phosphoproteomic studies, i.e., experiments are performed on different machines with different algorithms and search databases for searching, assembling, and quantifying from the raw MS spectra. In order to normalize the data quality in phosphopeptide identification, P3DB made a general data selection criterion with False Discovery Rate (FDR) less than 1 % and with less than 15 ppm precursor mass accuracy, which can be achieved by most of today's technologies.

The gene ontology terminologies are downloaded from the GO website [7, 8]. The proteome-wide annotations are imported from TAIR [9] and Phytozome [10]. The kinase and phosphatase family classification are consistent with PlantsP [11]. The protein information is also retrieved from PDB [12] and Uniprot [13]. The domain reference information is from Pfam A version 27.0 downloaded from the Pfam website [14]. The PPI data are collected from four online resources, including BioGRID [15], Intact [16], DIP [17], and MINT [18].

P3DB is an interactive user interface for online phosphorylation database with a back-end MySQL database, server-side PHP code, and front-end Javascript and jQuery libraries. Version 3.5 of P3DB complies with the standard HTML5 and CSS3. Jmol [19],

a java-based program, is used to visualize the protein tertiary structure. For network visualization and interactive analysis features, Cytoscape [20] JavaScript APIs are extensively used in P3DB, and they are compatible and reliable under most common browsers and operating systems.

3 Methods

Protein phosphorylation data can be browsed, queried, and visualized in the network framework in P3DB (Fig. 1). The methodology and procedures of targeting the data of interest for the corresponding phosphorylation events are described in this Subheading.

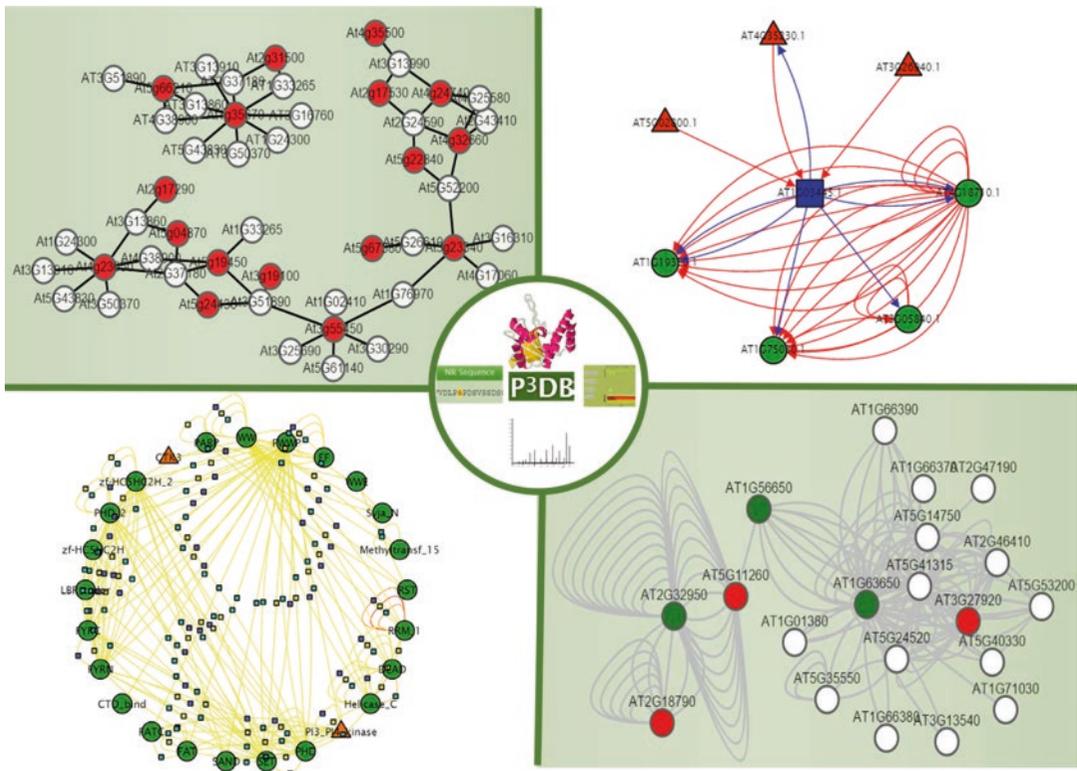


Fig. 1 Network tools in P3DB for the phosphorylation data analyses. Center: P3DB logo and phosphorylated protein and peptide information; *top left*: KiC-assay data visualization shows that a kinase and its substrates tend to form clusters; *top right*: a kinase-phosphatase-substrate network where a phosphatase could be activated by kinases and thereby play a central role in signaling; *bottom left*: a domain co-occurrence network connects the kinase domains, and WW, FF, PHD domains to illustrate the potential downstream gene regulation; *bottom right*: a protein-protein interaction network using the maximum linkage mode where the *right side cluster* is further expanded by its neighborhood

3.1 Explore and Browse the Data

P3DB provides multiple data browsers to explore the phosphorylation data sets, covered organisms, gene ontology, taxonomy, and kinase and phosphatase families. By exploring the data, the user can obtain a general idea of phosphorylation events and its linkage to possible functionalities across all the plant species. The curated plant phosphoproteomics data sets can be browsed directly under the browse tab (*see Note 1*). This general data browser organizes the data sets from the different studies categorized by species. While most of the studies contribute phosphoproteomics only in a single species, some of the studies contributed to multiple species using comparative phosphoproteomics. The throughput of each study varies due to the technology applied. For each study, the numbers of phosphorylated proteins and phosphosites are displayed. The summary statistics of the species are also displayed in a separate table. The organism-based data browser is suitable for comparative studies in phosphoproteomics (*see Note 2*). To facilitate large-scale data comparative analysis, P3DB provides APIs and data downloads. By taking advantage of P3DB data sets and performing in-depth mining, large-scale bioinformatics analysis and hypothesis generation on plant proteomics can be performed, like those studies in protein local disorder preference [21], sequence conservations for phosphorylation events [22], and conserved phosphorylation motifs [23].

The browser for gene ontology (available from the tool menu) is implemented in a tree structure, so that the user can narrow down the term or go up one level by clicking the gene ontology numbers. Three categories of gene ontology can be explored: biological process, molecular function, and cellular component. The left side box shows the gene ontology term hierarchy, and it always displays one level up and down. The right top box shows the detailed information for the specific gene ontology term, including name, definition, and synonyms. The right bottom box shows the proteins that are mapped to the gene ontology term, and if there is an active link, it points to a phosphoprotein in P3DB. This link directs the user to the phosphoprotein page where more information is displayed.

The taxonomy browser (available from the tool menu) is also designed in a tree structure. In P3DB, *Viridiplantae* is the top level taxonomy, and the data sets are displayed only at the species level in the leaf node. This provides a clear view for those who look for data sets of a specific species.

Kinase and phosphatase families can be browsed effectively in a smart table with hierarchical data structures (available from the browser menu). There are five main classes for kinases as well as one big class with three subclasses for phosphatase. It is noticeable that more than a thousand kinases and a hundred phosphatases are currently organized in this data structure.

The browsers in P3DB can lead to many interesting ideas and data analyses in plant phosphoproteomics. However, when a user is specifically interested in his or her own protein or peptide, querying within P3DB data sets provides better guidance and understanding of what has been discovered and what has not.

3.2 Make a Query

P3DB includes two types of search: sequence-based (BLAST) and text- or identifier- based search. In the former, BLAST is conducted against protein sequences with known phosphorylation sites in P3DB database. Then, the user can quickly learn if there are any existing studies or data related to the query that provide the most likely phosphorylation-related annotations (*see Note 3*). On the other hand, the BLAST function can also help users to find the whole set of homologous sequences for proteins or peptides across various species, which facilitates comparative phosphoproteomics studies.

Users can also query a protein by using the keywords of names or descriptions, or identification number. Moreover, since P3DB provides additional protein annotation data, users can also retrieve a protein by searching other features, such gene ontology annotations, keywords in kinase or phosphatase families, taxonomy, and protein domains. Users can target a class of proteins or complexes by searching against functional annotations, e.g., “ERK” or “malate transporter.”

If neither BLAST nor search gives any useful phosphorylation information, then the query protein or peptide is not yet covered in P3DB. P3DB then provides an alternative way to infer the phosphorylation events through machine learning-based prediction. Trained from the database high-quality experimental data sets, P3DB builds predictors for green plants and achieves state-of-the-art performance compared to other available plant phosphorylation prediction tools [24] (*see Note 4*).

The result of querying a protein or peptide leads to the phosphoprotein page. P3DB provides cross-links among protein pages, phosphosite pages, phosphopeptide pages, and spectra pages. It is convenient to browse the data in a top-down or bottom-up manner. P3DB compiles all of the existing studies as a whole so that the phosphoprotein, phosphopeptides, or phosphosites belonging to the same protein can be cross-referenced among multiple studies. This compiling of data sets provides a confidence of the phosphorylation events, and sometimes helps reveal the same or different phosphorylation patterns under different conditions (*see Note 5*). On the phosphoprotein page, many elements display protein annotations and their physicochemical properties, such as protein chart with hydrophobicity, domain and disorder scores, protein tertiary structure, a list of protein interactome, gene ontology annotations, kinase phosphatase assignments, protein domains, and even user comments.

3.3 PPI Network

In P3DB, the PPI data is available as a list on the phosphoprotein page. Alternatively, it can be searched and visualized in a network canvas. When the user provides a list of protein IDs in the search PPI text box (*see Note 6*) in the result page, the interactome data are first displayed in the smart table. The protein names, gene names, descriptions, and corresponding publications are shown in the table. The user can further filter and sort the records in the table before drawing the network. The PPI network is then displayed on the Cytoscape canvas after previewing the data. The nodes of the PPI network are proteins, and the edges show known interactions from the database. By clicking an edge, the detailed information for this pair of interaction is shown in a table, so that the data source of the interaction, the experiment type, and the original publication can be traced back if necessary.

The PPI network can be created in two modes: direct linkage and maximized connectivity. In the direct linkage mode, the network can be quickly created by loading the proteins in the query list and only searching interactions within these query proteins. The protein nodes sometimes form modules or islands in the network view. In such cases, the second mode is helpful to maximize the connectivity between each pair of nodes in the network. This requires calculation of the shortest path between each pair of the two protein nodes in the network, and then the network can guarantee the maximum connectivity between each of the two nodes. This mode can potentially uncover a protein interaction cascade. Although it may not reflect true signaling or metabolomics pathways, it could reveal potential protein complexes or functional relationships. Sometimes, the network forms modules that may represent individual functional groups. The user can further manually expand the network by exploring the neighborhood of the selected proteins (*see Note 7*). The nodes are color coded in the network, where red represents the requested proteins, green represents the proteins in the path of connections, and white shows the neighborhood nodes (bottom right box in Fig. 1).

In the process of exploring the PPI network, the protein phosphorylation events are also mapped onto the proteins. The potential correlation between a PPI network and phosphorylation can be explored case by case. The protein phosphorylation event itself can be viewed as a transient protein-protein interaction. Meanwhile, the phosphorylation usually changes the binding affinity of downstream protein-protein interactions.

3.4 Kinase Substrate and Phosphatase Substrate Network

Kinase-substrate and phosphatase-substrate networks are based on the known experimental phosphorylation or dephosphorylation events. These two networks can be overlaid on the same view. The input data needed to pull out the network is a list of protein identification numbers with each protein in one line. In the network view, the red triangles are the proteins annotated as kinases, while

the blue squares are the proteins designated as phosphatases. Substrate proteins are usually in green circles, while in some cases a substrate can also be a kinase or phosphatase. The phosphorylation event is marked with a red arrow, and the dephosphorylation event is marked with a blue arrow. Any nodes in the network can be selected and searched against the database to show the neighborhood centralized by the selected protein. In plant signaling pathways, the phosphorylation events are usually hierarchical, in which the user can see the kinase cascade in such a network (*see Note 8*). On the other hand, some phosphatases also need to be activated by phosphorylation through kinase interactions. These phenomena can be easily viewed in such a visualization framework (top right box in Fig. 1).

3.5 Protein Domain Co-Occurrence Network

Protein phosphorylation events are tied to functionalities of many types of protein domains, including kinase domains, phosphatase domains, activation domains (regulatory domains), and downstream binding domains (phosphorylation recognition domains). Studying the domain information is helpful to reveal phosphorylation mechanisms. Eukaryotic proteins, especially plant proteins, are mostly multiple domain proteins, where one protein has many consecutive domains. For example, protein kinases have both activation domains and kinase domains; the dual domain has both kinase and phosphatase roles; and many downstream transcription factors have both phosphorylation recognition domains and DNA binding domains. These consecutive domains in a single protein are called co-occurring domains. Co-occurring domains are not randomly interconnected, but usually have functional correlations. The domain co-occurrence network is a network consisting of nodes with domains, and edges indicating that the connected domains co-occur (bottom left box in Fig. 1).

In P3DB, a domain of interest can be searched by keywords, and the user can then select the correspondent domains to generate the domain network. The domain network is overlaid across multiple species, i.e., *Arabidopsis thaliana*, *Glycine max*, and *Oryza sativa*. The domains are marked with different colors and shapes for kinase (triangle triangle), phosphatase (light green square), and dual functions (purple hexagon), while other types of domains in regulation and downstream functions are marked in green circles. In a domain co-occurrence network, the domain information on each node and the protein information on each edge are displayed in the bottom boxes. The proteins that have phosphorylation events stored in P3DB are represented as links, which can redirect to the phosphoprotein pages. In this way, the user can see both phosphorylation events and the domain co-occurrence information so that the hypothesis of how phosphorylation regulates the functions and its insight molecular mechanism can be addressed to some degree. The user can also compare the domain compositions among different species (*see Note 9*).

4 Notes

1. Three menu items under the “browse” tab are directly targeted to the data sets, where “all” shows the whole list of phosphoproteins in P3DB; “data” lists the statistics by study indicating the contribution by organism, and “organism” provides the statistics by organism on phosphosites and phosphopeptides.
2. If users are interested in comparing phosphorylation data across different plant species, for phosphoproteins, phosphosites, or phosphopeptides, or to study sequence or function conservation, there are several ways to achieve this goal. For functions, a user can search the gene ontology terms or browse the terms to obtain all the phosphoproteins mapped to it. For specific proteins, the user can search by protein name or description, like “malate dehydrogenase,” and all of the homologous proteins can be obtained. The phosphorylation pattern can then be compared across different species. For sequence conservation, BLAST under the “search” protein functions is useful to get all of the conserved protein sequences or peptides based on homology, and then the phosphorylation pattern can be compared. Since these functions are in different modules (*see* Subheadings 3.1 and 3.2), we summarize them here.
3. BLAST function is available for both protein and peptide sequences. It is available in “Protein” option under the “Search” menu. There are three parameters for BLAST function. The user can select the organism and the data sets to perform the search. The E-value cutoff can also be specified by the user.
4. For the phosphorylation site prediction, the user can choose to predict Ser/Thr phosphorylation or Tyr phosphorylation by selecting the proper models. The result page will be linked to the Musite web services [25]. In the Musite website, the user is allowed to set more parameters. Musite is a comprehensive protein post-translational modification predictor, which provides more models to choose from and it is not limited to plants.
5. On the phosphoprotein, phosphopeptide, or phosphosite page, there is a dropdown list to show all of the studies reporting this phosphorylation event. On the phosphoprotein page, the phosphorylation sites are overlaid with multiple studies, and the user can filter with a specific study to compare phosphorylation patterns. This is useful when some phosphorylation sites are supported by many experiments. This provides some idea about the power of conservation of phosphorylation events.
6. The input of the PPI network search is a list of protein identification numbers with a space or comma as the separator. If any protein identification number is not in the P3DB database, the system will ignore this protein automatically, and the search will not cause an error.

7. The PPI network is under “search” and “PPI” or “tool” and “PPI.” The Cytoscape view provides many flexibilities in the display. The network can be zoomed in or out and relocated by controlling the wheel and bar on the left side of the graph. It also provides many different layouts to display the network. While arbor is a default layout, grid is useful when there are too many nodes in the network, and the circle layout is useful when there are too many interactions. The buttons above the network can be used to tune the network further.
8. In order to view the kinase cascade in plant biology, a user can select the kinase or protein as a substrate and click “search by selected.” The network is then regenerated by the protein selection and centered on it. The network view can also be saved in high quality for preview and publication purposes.
9. In a domain co-occurrence network, the user can filter the network by clicking on the species names. Also, the edges are marked with a color gradient based on the number of proteins mapped to the co-occurring domains. The domain network can be viewed in a full screen and also saved as a high-quality image.

Acknowledgments

This work was supported by funding from the National Institute of Health (Grants GM078601 and GM100701) and the National Science Foundation (Grant DBI-0604439). We would also like to express thanks to Drs. Jay Thelen and Jianjiong Gao for their helpful input.

References

1. Huber SC (2007) Exploring the role of protein phosphorylation in plants: from signalling to metabolism. *Biochem Soc Trans* 35(Pt 1): 28–32. doi:[10.1042/BST0350028](https://doi.org/10.1042/BST0350028)
2. Park CJ, Caddell DF, Ronald PC (2012) Protein phosphorylation in plant immunity: insights into the regulation of pattern recognition receptor-mediated signaling. *Front Plant Sci* 3:177. doi:[10.3389/fpls.2012.00177](https://doi.org/10.3389/fpls.2012.00177)
3. Heazlewood JL, Durek P, Hummel J, Selbig J, Weckwerth W, Walther D, Schulze WX (2008) PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res* 36(Database issue):D1015–D1021. doi:[10.1093/nar/gkm812](https://doi.org/10.1093/nar/gkm812)
4. Yao Q, Ge H, Wu S, Zhang N, Chen W, Xu C, Gao J, Thelen JJ, Xu D (2014) P(3)DB 3.0: from plant phosphorylation sites to protein networks. *Nucleic Acids Res* 42(Database issue):D1206–D1213. doi:[10.1093/nar/gkt1135](https://doi.org/10.1093/nar/gkt1135)
5. Huang Y, Thelen JJ (2012) KiC assay: a quantitative mass spectrometry-based approach for kinase client screening and activity analysis [corrected]. *Methods Mol Biol* 893:359–370. doi:[10.1007/978-1-61779-885-6_22](https://doi.org/10.1007/978-1-61779-885-6_22)
6. Ahsan N, Huang Y, Tovar-Mendez A, Swatek KN, Zhang J, Miernyk JA, Xu D, Thelen JJ (2013) A versatile mass spectrometry-based method to both identify kinase client-relationships and characterize signaling network topology. *J Proteome Res* 12(2):937–948. doi:[10.1021/pr3009995](https://doi.org/10.1021/pr3009995)
7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese

- JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25–29. doi:[10.1038/75556](https://doi.org/10.1038/75556)
8. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res* 43(Database issue):D1049–D1056. doi:[10.1093/nar/gku1179](https://doi.org/10.1093/nar/gku1179)
 9. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202–D1210. doi:[10.1093/nar/gkr1090](https://doi.org/10.1093/nar/gkr1090)
 10. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40(Database issue):D1178–D1186. doi:[10.1093/nar/gkr944](https://doi.org/10.1093/nar/gkr944)
 11. Gribskov M, Fana F, Harper J, Hope DA, Harmon AC, Smith DW, Tax FE, Zhang G (2001) PlantsP: a functional genomics database for plant phosphorylation. *Nucleic Acids Res* 29(1):111–113
 12. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* 39(Database issue):D392–D401. doi:[10.1093/nar/gkq1021](https://doi.org/10.1093/nar/gkq1021)
 13. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol Biol* 1374:23–54. doi:[10.1007/978-1-4939-3167-5_2](https://doi.org/10.1007/978-1-4939-3167-5_2)
 14. Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JÉ, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211–D222. doi:[10.1093/nar/gkp985](https://doi.org/10.1093/nar/gkp985)
 15. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 41(Database issue):D816–D823. doi:[10.1093/nar/gks1158](https://doi.org/10.1093/nar/gks1158)
 16. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeifferberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 40(Database issue):D841–D846. doi:[10.1093/nar/gkr1088](https://doi.org/10.1093/nar/gkr1088)
 17. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32(Database issue):D449–D451. doi:[10.1093/nar/gkh086](https://doi.org/10.1093/nar/gkh086)
 18. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40(Database issue):D857–D861. doi:[10.1093/nar/gkr930](https://doi.org/10.1093/nar/gkr930)
 19. Herraiz A (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem Mol Biol Educ* 34(4):255–261. doi:[10.1002/bmb.2006.494034042644](https://doi.org/10.1002/bmb.2006.494034042644)
 20. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3):431–432. doi:[10.1093/bioinformatics/btq675](https://doi.org/10.1093/bioinformatics/btq675)
 21. Yao Q, Bollinger C, Gao J, Xu D, Thelen JJ (2012) P(3)DB: an integrated database for plant protein phosphorylation. *Front Plant Sci* 3:206. doi:[10.3389/fpls.2012.00206](https://doi.org/10.3389/fpls.2012.00206)
 22. Yao Q, Gao J, Bollinger C, Thelen JJ, Xu D (2012) Predicting and analyzing protein phosphorylation sites in plants using musite. *Front Plant Sci* 3:186. doi:[10.3389/fpls.2012.00186](https://doi.org/10.3389/fpls.2012.00186)
 23. van Wijk KJ, Friso G, Walther D, Schulze WX (2014) Meta-analysis of Arabidopsis thaliana phospho-proteomics data reveals compartmentalization of phosphorylation motifs. *Plant Cell* 26(6):2367–2389. doi:[10.1105/tpc.114.125815](https://doi.org/10.1105/tpc.114.125815)
 24. Yao Q, Schulze WX, Xu D (2015) Phosphorylation site prediction in plants. *Methods Mol Biol* 1306:217–228. doi:[10.1007/978-1-4939-2648-0_17](https://doi.org/10.1007/978-1-4939-2648-0_17)
 25. Gao J, Thelen JJ, Dunker AK, Xu D (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics* 9(12):2586–2600. doi:[10.1074/mcp.M110.001388](https://doi.org/10.1074/mcp.M110.001388)

Navigating the Glycome Space and Connecting the Glycoproteome

Matthew P. Campbell, Robyn A. Peterson, Elisabeth Gasteiger, Julien Mariethoz, Frederique Lisacek, and Nicolle H. Packer

Abstract

UniCarbKB (<http://unicarbkb.org>) is a comprehensive resource for mammalian glycoprotein and annotation data. In particular, the database provides information on the oligosaccharides characterized from a glycoprotein at either the global or site-specific level. This evidence is accumulated from a peer-reviewed and manually curated collection of information on oligosaccharides derived from membrane and secreted glycoproteins purified from biological fluids and/or tissues. This information is further supplemented with experimental method descriptions that summarize important sample preparation and analytical strategies. A new release of UniCarbKB is published every three months, each includes a collection of curated data and improvements to database functionality. In this Chapter, we outline the objectives of UniCarbKB, and describe a selection of step-by-step workflows for navigating the information available. We also provide a short description of web services available and future plans for improving data access. The information presented in this Chapter supplements content available in our knowledgebase including regular updates on interface improvements, new features, and revisions to the database content (<http://confluence.unicarbkb.org>).

Key words Curation, Database, Glycan, Glycobioinformatics, Glycomics, Glycoproteomics

1 Introduction

Advances in glycomics and glycobiology research have led to a wealth of accumulated data with the exciting potential to richly inform our understanding of the biological implications of glycosylation. However, sustainable bioinformatics tools are essential for data and information storage to maximize the value of research efforts [1–3]. UniCarbKB [4, 5] represents a continued effort to build upon the success of EUROCarbDB [6] and GlycoSuiteDB [7], to provide a sustainable framework that supports the growth of glycobioinformatics and the dissemination of knowledge through the provision of an open and unified portal. Ultimately,

UniCarbKB encourages the sharing and cross-referencing of datasets, thereby making glycomics data more accessible and discoverable to the general life science community.

In 2012, at its launch, UniCarbKB was seeded with curated glycan and glycoprotein data sourced from GlycoSuiteDB and the EUROCarbDB initiatives. Over the last few years the content and diversity of UniCarbKB has grown as a result of our continuing efforts to curate published data spanning the last two decades, with a focus on mammalian glycoproteins and their associated glycoforms, often accompanied with known sites of attachment.

Our biocuration aspirations have focused on delivering a quality glycan and glycoprotein hub of information—entries are continuously reviewed to keep up with current scientific literature—that, in collaboration with UniProtKB/Swiss-Prot [8] provides a gold standard set of experimentally verified data. As part of our curation program, we are constructing a controlled vocabulary and ontology for describing the captured structural and experimental data that adheres to the principles of the MIRAGE [9] and the GlycoRDF [10] initiatives to improve data interoperability. Protein names use standardized nomenclature and synonyms from the literature and other databases (e.g. UniProtKB/Swiss-Prot), with associated glycosyltransferase enzyme-specific information, tissue localization, and species of origin, accumulating in 4000 structures characterized from over 850 glycoproteins, an additional 160 glycoproteins with compositional only data are available, sourced from 910 publications. Approximately 580 entries are cross-referenced with UniProtKB.

The knowledgebase offers a freely accessible and information-rich resource supported by querying interfaces, annotation technologies, and the adoption of common standards to integrate structural, experimental, and functional data.

In this chapter, we outline a selection of step-by-step protocols for searching and navigating UniCarbKB. These protocols are representative of typical queries and methods performed by researchers to find information from quality published research on individual glycan structures, attached glycoproteins (at the global and site-specific level), and specific metadata including experimental methods and validation techniques sourced from individually curated publications. Users can choose to search UniCarbKB by (sub)structure, monosaccharide composition, glycan mass, taxonomy, tissue, glycoprotein (accession number or UniProtKB/Swiss-Prot name) or by literature publication.

2 Materials

UniCarbKB is freely accessible and hosted at <http://www.uni-carbkb.org/>. This Chapter will focus on all entries, tools, and resources available in the UniCarbKB database. The database is

updated every three months with each new release including provision of newly curated material, edits to existing entries, and feature updates as a result of improving technologies and user feedback. Recently, we have focused on adding information for mammalian glycoproteins absent from our existing library to improve coverage of the glycoproteome. Typically, we select publications that: (1) report fully characterized glycan structures with minimal linkage and monosaccharide ambiguity; (2) use analytical methodologies including mass spectrometry, liquid chromatography and capillary electrophoresis (with exoglycosidase digestions), and nuclear magnetic resonance; and (3) provide site-specific data on amino acid positions or peptide sequences that are in agreement with indicated glycosylation sites in UniProtKB, with validated *N*-linked glycans. Large-scale/high-throughput analysis of glycoproteins isolated from a biological source is supported by UniCarbKB, but such glycan data are typically compositional only with assignments inferred from known biosynthetic rules.

2.1 Implementation

UniCarbKB is built with the open-source Play Framework (<http://www.playframework.com>) written in Java and Scala. The user-interface is predominantly written in Scala and the front-end is partially built with the popular Bootstrap framework. The model and controller layers are written in Java and we use the Ebean object-relational mapping library to query the PostgreSQL database model.

2.2 Databases Linked to UniCarbKB

A key objective of UniCarbKB is to forge relationships with externally hosted structural and experimental glycan databases. Specialized information within the scope of UniCarbKB is made available via cross-references to other relevant databases, such as the PubChem [11], NCBI Chemical database; SugarBindDB [12, 13], the pathogen–glycan interaction database; UniCarb-DB [14], a MS/MS experimental database; and GlycoMob [15], an Ion Mobility Collision Cross Section database.

Efforts to cross-reference UniCarbKB with SugarBindDB, the universal glycan repository GlyTouCan [16] and the molecular modeling platform GLYCAM [17] are on-going. For example, glycan ligands of bacteria stored in SugarBindDB are systematically matched to UniCarbKB (via the GlyS3 substructure search [18]) and each matching ligand entry is cross-linked to corresponding UniCarbKB entries.

One of the strengths of UniCarbKB is the connectivity with UniProtKB/Swiss-Prot—a detailed description of this partnership is described in Subheading 3.9. The GlycoMod tool [19] (<http://web.expasy.org/glycomod>) can be used to predict oligosaccharide structures from experimentally determined masses and is directly linked to UniCarbKB, connecting theoretically possible compositions with curated glycan structures and thereby helping the user interpret their prediction results.

For all glycan structure entry and publication summary pages (*see* below), implicit links to these affiliated databases are listed under the “Databases” tab. A list of cross-referenced databases and the current state of development is available at <http://www.unicarbkb.org/crossreferences>, including the forthcoming release of UniCorn, a theoretical database of *N*-glycan structures.

2.3 Data Availability—Web Services

UniCarbKB provides programmatic access to a range of services via web service interfaces based on REST (Representational State Transfer). The availability of web services allows the integration of data into other tools, applications, and workflows. Currently, the web services are limited to structure-based searches, and supporting documentation is provided via a Swagger-enabled API at <http://www.unicarbkb.org/rest-api>. A detailed explanation of the web services will be published elsewhere. However, through these services users can search UniCarbKB using common programming languages to retrieve structure-associated data. For example, users can retrieve structure identifiers by searching for an exact structure. Part of these services are implemented by the GlycoPattern database [20], which allows for sharing and cross-referencing of structure entries.

3 Methods

3.1 Using the UniCarbKB Website

The UniCarbKB team has benefited from regular user feedback and analysis of user activity on the site, which has been instrumental in developing intuitive querying/search tools. To retain a sense of familiarity across glyco-related databases, the layout and functionality of UniCarbKB borrows components from GlycoSuiteDB and EUROCarbDB. An emphasis has been placed on supporting the most frequently used functionalities. For example, database searches with simple queries (only a few terms) have been enhanced with native selects by integrating an auto-complete suggestion mechanism. Presentation and navigation of medium- to large-sized results has also been improved with extended use of pagination across the website. With the continuing growth and diversity of UniCarbKB, we have strived to improve sections of the user-interface to simplify content layout. To this end, viewing a database entry has been significantly improved with simplified terminology, removal of redundant descriptions and the availability of documentation/troubleshooting guides.

3.2 The UniCarbKB Home Page

The database homepage, accessible at <http://www.unicarbkb.org>, is the main entry point for most users (*see* **Note 1**). The home page consists of four main sections:

1. A navigation bar with drop-down menus and links providing direct access to the data and functionality available in the UniCarbKB website.
2. A brief summary introducing the objectives and role of UniCarbKB including database statistics, the participating expert research institutions, and our funding sources.
3. Links to affiliated database efforts including UniCarb-DB, GlycoMob, SugarBindDB, and UniCorn plus international activities such as GLYCAM, GlyTouCan, and GlycoBase.

3.3 UniCarbKB Navigation Banner

Some elements are always present at the top of the page: the UniCarbKB logo to return to the homepage, access to structure and biological source search tools, links to curated references, external databases, data interpretation tools, help guides, and contact information (Fig. 1). Further information for each listing is described in the following Subheadings. For information on resources available from the “Tools” link including GlycoDigest and GlyS3, the sub-structure search tool readers are referred to [18, 21].

3.4 Search Functionalities

Direct access to the search tools is available from the “Structure” and “Biological Source” menus. Here, the “Structure” menu partitions structural search features into composition, GlycoCT, and

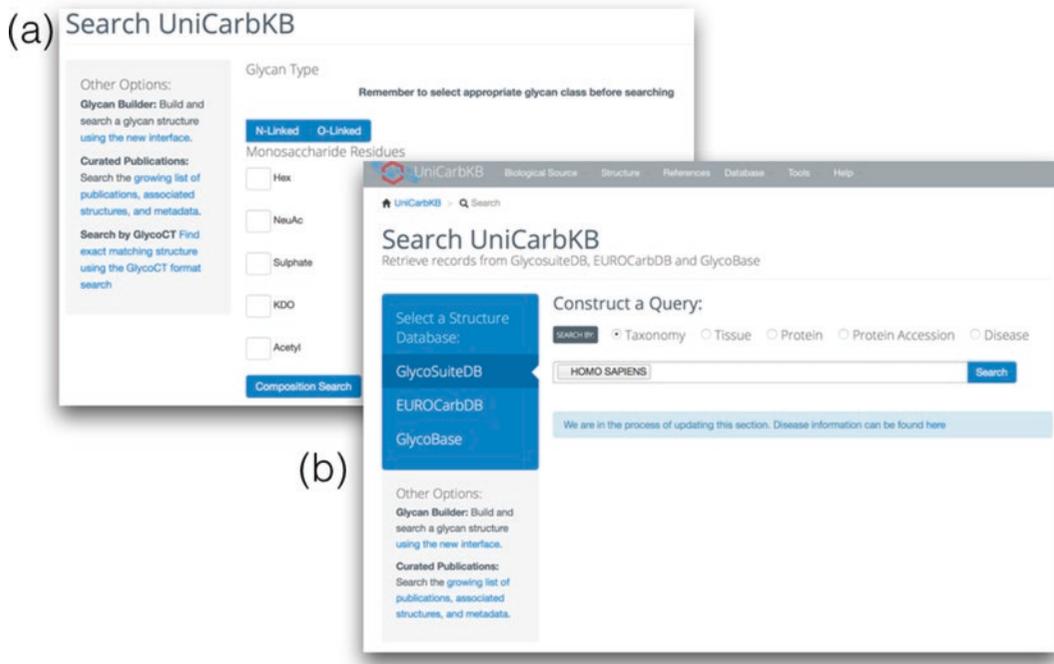


Fig. 1 Users can search the database content by (a) monosaccharide composition or (b) attached protein, taxonomy, or tissue. To access these search forms, UniCarbKB selects either the query options from the Structure or Biological Source menus on the top navigation bar

graphical. In contrast, “Biological Source” encapsulates taxonomic, tissue and glycoprotein (accessible from the “Query Options” link) and cell line searches (*see* **Note 2** and Fig. 1).

3.4.1 Searching for Information Associated with Taxonomy, Tissue, or Glycoprotein

The “Biological Source” search page has search options and data filters available. In this Subheading, we briefly describe the basic queries available and their usage:

The “Biological Source” menu, accessible from the navigation bar, lists four search items: “Taxonomy”, “Tissue”, “Protein”, and “Protein Accession”

1. To perform a query, select a preferred query option—this will automatically load a search text box.
2. After selecting the appropriate search option start typing a name or keyword(s) into the search box at the top of the query screen.
3. After entering three characters, the quick navigation aid automatically provides a dropdown list of matching items, matched by NCBI Taxon, MeSH (Medical Subject Headings) or protein name (*see* **Note 3**).
4. The search box supports the selection of multiple items from the previous step.
5. Pressing the search button will load data available for the selected item(s).

Results for the specified taxonomy, glycoprotein, or tissue search will appear on a new screen with the following information:

1. An icon representing the content type (taxonomy, tissue, or glycoprotein) with the title/name of the content search, which is linked to the item’s main viewing template.
2. A summary of the most relevant few lines of content related to the item. In brief, taxonomy searches list a limited number of glycoproteins; for the tissue and glycoprotein results only the number of structures is shown.
3. An icon indicating the number of structures is shown to the right of each text block.

3.4.2 Taxonomy

In addition to structure and glycoprotein data, UniCarbKB contains taxonomic terms, based on the NCBI taxonomy database. NCBI Taxonomy IDs are assigned to each species entry in the database. A similar approach was used by EUROCarbDB, GlycoBase [22] and continues to be used by Glycome-DB [23], SugarBindDB and the Carbohydrate Structure Database [24] amongst others, thus permitting interoperability between these glyco-focused resources.

Taxon searches are case-insensitive and to assist users, an auto-complete feature is available. The search bar supports the selection of multiple taxonomies from the list of matching entries, simply by searching and selecting a species of interest (*see* **Note 4**).

1. To find content associated with the species *Homo sapiens*, select the “Query Options” link from the “Biological Source” menu, which can be found in the top navigation bar.
2. Select the “Taxonomy” button in the query page and enter *Homo sapiens*. Species partially matching the text entered shall be listed automatically. Once you have selected *Homo sapiens* press the search button.
3. The results page will display the search term and the number of structures (including reported compositions) along with a short summary of glycoproteins. To continue browsing data available for *Homo sapiens*, select the blue structures button.
4. On the summary page, the left-column lists all glycan structures and compositions in UniCarbKB associated with *Homo sapiens*. By default, only ten glycan structures are displayed—to view the complete catalog of structures click the “Show all structures” button (*see Note 5*). Since we have restricted the result listing to a taxonomy search, the right-column sidebar will only list relevant “Biological Associations”; for example, all glycoproteins characterized from the species *Homo sapiens*.
5. To view all glycoproteins, click the “Biological Associations” Proteins label. Users can navigate to the glycoprotein page by scrolling the list to find any protein of interest.

3.4.3 Tissue

To ensure consistency across UniCarbKB, the actions for searching tissue data are similar to the above taxonomy workflow. Similar to other databases, UniCarbKB adopts the NCBI MeSH (Medical Subject Headings) vocabulary for tissue annotations, whereby each tissue or biological source entry is linked with a NCBI MeSH identifier. Searches are case-insensitive and the search bar provides suggestions after entering three characters.

3.5 Glycan Structure Searching

Glycans are inherently complex biomolecules, involving a breadth of monosaccharide diversity and varying degrees of branching. Consequently, a simple textual representation that encodes monosaccharide anomericity, glycosidic linkages, residues modifications and substitutions, and structure ambiguity is difficult to achieve. A plethora of graphical and textual formats are used for the depiction of glycan structures including: Consortium for Functional Glycomics/Essentials, Oxford nomenclature, and IUPAC format [25–28]. A combination of these formats in which the linkages are depicted as angles as in the Oxford scheme, on the colored residue symbols of the “Essentials in Glycobiology” scheme, is gaining acceptance and is supported by the UniCarbKB suite of tools and databases. Recently, editors of the Third Edition of “Essentials in Glycobiology” announced an extended monosaccharide library that adopts the CFG/Essentials color notation with optional Oxford depiction of monosaccharide linkages with embedded specificity and anomericity [29].

In addition to the graphical representation of oligosaccharides, a variety of computational encoding formats have been developed notably GlycoCT [30], LinearCode® [31], KCF (KEGG Chemical Function) [32], LINUCS [33], and WURCS [34]. A comprehensive review and comparison of nomenclatures have been previously published [35].

To find a structure of interest in UniCarbKB, users can either (1) use GlycanBuilder, the GlycoCT form or the composition query engine, or (2) use the contextual navigation options presented in the viewing templates (for more information refer to individual page descriptions below).

3.5.1 Using GlycanBuilder to Perform (Sub)Structure Searches

GlycanBuilder is a versatile tool that allows for the fast and intuitive drawing of glycan structures that can be used standalone, embedded in web pages or integrated into other programs [36, 37]. The tool was developed during the EUROCarbDB project in conjunction with the mass spectrometry data interpretation software GlycoWorkBench. An updated version of GlycanBuilder (<http://bit.ly/1FwRxbo>) is included in UniCarbKB to support structure queries.

GlycanBuilder, accessed at <http://www.unicarbk.org/builder>, can be used to (1) build a new glycan structure, (2) build a substructure or epitope, or (3) extend a structure by using the library of predefined structures. Users are encouraged to refer to guides and documentation provided by the developers (<http://bit.ly/24heTlh> and <http://bit.ly/1rqBIBe>). In brief, the tool comprises three main sections: (1) drawing canvas, (2) monosaccharide and linkage panel, and (3) menu options for exporting images, switching between symbol/notation formats and selecting monosaccharide residues. The tool provides a selection of starting structures, e.g. *O*-glycan cores and hybrid type *N*-links to help users get started.

3.5.2 Query by Structure

Upon submitting a structure query, the default setting retrieves database entries that have exact matching topology, linkage, and anomeric configuration. A level of fuzziness is allowed by partial structure searching, whereby unknown information is handled as wildcards by the search algorithm. For substructure searching, only structures matching the epitope or extended motif will be listed in the results page.

3.6 Glycan Structure Summary Page

Glycan structure and its supporting evidence is a central focus of UniCarbKB. The glycan structure page is the place where all information connected with a structure can be easily accessed. Found here is a standard set of data types, and an overview of content available. This subheading leads users through the basics of navigating content associated with a specific glycan entry; an example screenshot is shown in Fig. 2.

UniCarbKB Biological Source Structure References Database Tools Help

UniCarbKB > Structure (UC 4027)

Structure

Details for this N-LINKED glycan structure

Structure Format

CFG/ESSENTIALS TEXT OXFORD

Linkage Type

N-LINKED glycan structure

Biological Associations

TAXONOMY (26) PROTEIN (13) SOURCE (14)

Defined Content

SOURCE (14)

External Links

PUBCHEM ENTRY

GlycoDigest

SIMULATE ENDOGLYCOSIDASE DIGESTION

References associated to structure

Title	Year	Authors
Structural characterization of the N-glycan moiety and site of glycosylation in vitellogenin from the decapod crustacean <i>Cherax quadricarinatus</i> .	2004	Khalaila I, Peter-Katalinic J, Tsang C, Radcliffe CM, Afialo ED, Harvey DJ, Dwek RA, Rudd PM, Sagi A
Tandem mass spectrometry of ribonuclease A and B: N-linked glycosylation site analysis of whole protein ions	2002	Reid, Stephenson, McLuckey
N-linked oligosaccharides of cobra venom factor contain novel alpha(1-3)galactosylated Le(x) structures.	2001	Gowda DC, Giushka J, van-Halbeek H, Thotakura RN, Bredehorst R, Vogel C-W
Analysis of Asn-linked glycans from vegetable foodstuffs: widespread occurrence of Lewis a, core alpha1,3-linked fucose and xylose substitutions.	2001	Wilson IB, Zeleny R, Kolarich D, Staudacher E, Stroop CJ, Kamerling JP, Altmann F
Characterization of the glycosylation sites in cyclooxygenase-2 using mass spectrometry.	2001	Nemeth J, Hochgesang G, Marnett L, Caprioli R, Hochsensang G
On-target endoglycosidase digestion matrix-assisted laser desorption/ionisation mass spectrometry of glycopeptides	2001	Colangelo, Orlando

Fig. 2 Each structure entry in UniCarbKB has three parts: (1) structure listing (*left-pane*) displaying the structure (in a defined nomenclature format) with linkage and monosaccharide anomericity depicted along with a description of supporting references; (2) the “Biological Associations” subsection, which summarizes the glycan distribution in the context of taxonomies, tissues and glycoproteins; and (3) links to external database including PubChem

A typical glycan structure page is divided into four main sections:

1. Structure image—on the left side, the structure is displayed using the Essentials/CFG nomenclature as default.
2. Supporting publications—a summary table is provided of the published literature curated in UniCarbKB that contains experimental evidence for the glycan structure. To view any of the publications in detail click the title to open the Reference page.
3. Glycosyltransferases (limited to *N*-glycans)—if the structure is fully defined with no sequence fuzziness or linkage ambiguity, an enzyme tab appears. This feature, called GlycanSynth, indicates the enzyme glycosyltransferases (GTs) involved in the biosynthesis of each disaccharide unit. A list of these gene names, enzymes, and reactions is available at <http://unicarbk.org/enzymes>.

4. Sidebar—displays biological association metadata, links to external databases, calculated mass and structure classification. The biological association submenu concisely groups taxonomic, glycoprotein, and tissue source information pertinent to the displayed structure (refer to Subheading 3.7.2).

3.6.1 Switching Between Symbol Nomenclatures

Symbol notation provides a compact yet visual approach for describing complex glycans and contextually provides an efficient, easier to understand means for annotating complex datasets. By default, the hybrid CFG/Essentials with linkage format is set for displaying structures in UniCarbKB. To convert between the “CFG/Essentials”, “Oxford”, and “Traditional” formats, users can select an option in the “Structure Format” sidebar panel.

3.6.2 Glycan Search by Composition

Drawing glycan structures and defining glycosidic linkages can be a bit cumbersome. To help users perform more efficient searches, a composition search tool is available (<http://www.unicarbkb.org/structureQuery>). The composition search form comprises 14 boxes corresponding to individual monosaccharide residues and modifications. By default, each composition is set to zero, and treated as underivatized. To perform a composition search:

1. From any page select the “Structure” tab from the navigation bar (top of the page) followed by the “Composition” option. You will be presented with the search page.
2. Select the *N*-linked or *O*-linked button to choose searches of a specified glycan class (*see Note 6*).
3. Enter whole integer values into any of the monosaccharide composition boxes and click the Search button.
4. Each glycan structure is linked to a summary page, which can be accessed by clicking the image of interest.
5. To narrow the results table, the “Filter” button can be used to show entries associated with a specific glycoprotein or taxon name.

The results page, in a simple table, lists those structures matching the exact composition only (*see Note 7*). Where known, additional information summarizes glycoproteins bearing each structure alongside the relevant species. Notable features include: the queried composition is displayed in line with the page header; and akin to many pages, the “Structure Format” button enables users to convert between supported nomenclature formats.

3.7 Publication Lists—Curated Content

The UniCarbKB offers the community access to a growing, curated database of information on the glycan structures of glycoproteins. Our biocuration activities are focused on publications describing the detailed characterization of mammalian glycoproteins, however,

with growth in large-scale glycoproteomics we are steadily improving the integration of site-specific glycosylation data acquired from such studies.

A complete listing of publications is available at <http://www.unicarbk.org/references> or via the “References” tab (navigation bar). The summary table provides a snapshot of the number of structure entries sourced from the publication in addition to title, year of publication, authors, and journal name metadata. Here, pagination is used to improve content layout and provides a solution for displaying a large data collection.

Users can navigate the paginated content either by (1) scanning through the list using the page counters or (2) by using the filter box to match an author name or publication title. For example, entering cancer will retrieve publication records with the word “cancer” in the title; alternatively typing an author name will list all appropriate papers.

1. On the home page, select the References tab from the navigation bar at the top of the page.
2. Use the filter bar to retrieve publication title(s) or author(s) partially matching the specified search term.
3. To view individual publication content including glycan structures, biological context and experimental information, users can click the publication title.

3.7.1 Individual Publications

As shown in Fig. 3, the publication template is divided into four distinct sections:

1. Header displaying bibliographic information.
2. Abstract sourced and linked to PubMed.
3. As shown in the Structure section, a graphical listing of glycan structures curated from the publication, which are linked to the “Glycan Structure” page.
4. The sidebar provides contextual navigation including “Biological Associations”, “Validation Method”, and “Connections”. Each of these subsections is described in the following Subheadings.

3.7.2 Biological Associations

The (individual or multiple) taxonomic, glycoprotein, and tissue sources of a set of glycan structures can be found under “Biological Associations”. Each classification is grouped into expandable drop-down lists, as shown in Fig. 2. To find related biological content you can use the embedded links in the quick navigation lists to access individual taxonomy, glycoprotein, or tissue source pages.

3.7.3 Validation Methods

This subheading details the experimental methods, techniques, and instrumentation reported in the corresponding publication, including (1) sample preparation and glycan release techniques exemplified but not limited to exoglycosidase treatment, derivatization, or

UniCarbKB Biological Source Structure References Database Tools Help

UniCarbKB > References > Human sperm binding is mediated by the sialyl-Lewis(x) oligosaccharide on the zona pellucida, 2011, "SCIENCE"

"SCIENCE", 2011

Reference details CURATED ENTRY GLYCOSUITEDB

Pang PC, Chiu PC, Lee CL, Chang LY, Panico M, Morris HR, Haslam SM, Khoo KH, Clark GF, Yeung WS, Dell A.

Human fertilization begins when spermatozoa bind to the extracellular matrix coating of the oocyte, known as the zona pellucida (ZP). One spermatozoan then penetrates this matrix and fuses with the egg cell, generating a zygote. Although carbohydrate sequences on the ZP have been implicated in sperm binding, the nature of the ligand was unknown. Here, ultrasensitive mass spectrometric analyses revealed that the sialyl-Lewis(x) sequence NeuAcα2-3Galβ1-4(Fuca1-3)GlcNAc, a well-known selectin ligand, is the most abundant terminal sequence on the N- and O-glycans of human ZP. Sperm-ZP binding was largely inhibited by glycoconjugates terminated with sialyl-Lewis(x) sequences or by antibodies directed against this sequence. Thus, the sialyl-Lewis(x) sequence represents the major carbohydrate ligand for human sperm-egg binding.

PubMed Entry: [21852454](#)

Structure Format
CFG/ESSENTIALS TEXT OXFORD

Biological Associations
TAXONOMY PROTEIN SOURCE

Validation Method
 PNGASE F
 PERMETHYLATION
 ALPHA 2,3 NEURAMINIDASE
 MALD-TOF-TOF
 CID/CAD
 PERMETHYLATION WITH NAOH/DIMETHYL SULFOXIDE

Fig. 3 Publication page for an article by Pang et al. reporting oligosaccharides characterized from human zona pellucida. The top content panel includes the title of the publication and journal details. The main body (*left-hand panel*) shows the abstract and link to PubMed along with a list of experimentally validated structures, which are linked to the structure summary page. For each publication, we record the following details under the “Validation Methods” subsection in the right-hand panel: (1) sample preparation procedures and glycan release techniques and/or methods that alter glycan structure, including exoglycosidase treatment and derivatization; (2) the analytical (mass, sequencing, and linkage) approach; and (3) supporting validation methods. In addition, grouped in the “Biological Associations” subsection we capture species, tissue, and glycoprotein(s) information

esterification of sialic acids; (2) the analytical strategies employed to determine the reported structural, e.g. mass, sequencing, and linkage; and (3) complementary validation methods that may include lectins/glycan binding proteins. The goal is to align the validation metadata terms with the objectives of MIRAGE, i.e. to provide sufficient evidence for researchers to assess the reliability, specifics, and accuracy of the information presented.

3.8 More About the Sidebar

The sidebar features page-specific links which improve navigation and the logical organization of UniCarbKB. The sidebar appears on a majority of pages, except for the References and Glycoprotein summaries (*see Note 8*). By using the sidebar you can:

1. Choose a specific “Structure Format” for glycan display.
2. Perform various operations via the links on the sidebar.

3. Use the contextual navigation options that appear under each Subheading, based on the type of content on the page; for example, investigate relevant Biological Associations.
4. Access external databases and supporting tools.

The links in this area change depending on the section of database you are viewing. The sidebar provides important links to internal and external pages.

3.9 Glycoprotein Searches and Links with UniProtKB

There have been few intensive programs to cross-reference glycan-related databases with those that support genomics and proteomics research. Previously, GlycoSuiteDB was the sole database linked from the UniProt Knowledgebase for additional curated glycoprotein information, and for submitting some annotations directly into UniProtKB/Swiss-Prot. Following the launch of UniCarbKB, a two-way program has been established to share new glycoprotein knowledge. In May 2013 UniProtKB switched cross-reference links to UniCarbKB instead of (the no longer maintained) GlycoSuiteDB (UniProt release 2013_06), such that all UniProtKB glycosylation entries associated with GlycoSuiteDB were updated to link to UniCarbKB and vice versa. Thanks to these cross-links, UniProtKB users have direct access to structural information and corresponding meta-data for characterized glycoproteins, which often exceeds the scope of UniProtKB as a generalist protein resource.

The collaboration established between UniCarbKB and UniProtKB goes beyond mere cross-referencing. Experimentally determined glycosylation sites curated in UniCarbKB, along with their references in the scientific literature, are submitted to UniProtKB on a regular basis, where the data are reviewed by senior biocurators before inclusion in UniProtKB/Swiss-Prot protein entries.

On the other hand, UniCarbKB protein descriptions are based on the preferred name and primary accession identifier denoted by UniProtKB [8]. However, in cases where UniProtKB does not describe the protein, or where the protein cannot be clearly identified in UniProtKB, we use the protein name described in the cited publication. In addition, UniCarbKB glycoprotein summary pages are enriched by dynamically including PTM-related annotations and protein sequences from UniProtKB/Swiss-Prot. An example of this integration between the two databases can be seen in Fig. 4, Subheading 3.10.1.

Although the number of fully characterized glycoproteins is limited at this stage, for each glycoprotein record, two levels of annotation are provided where known: (1) global-specific data denote all glycan structures characterized on a single purified glycoprotein; (2) site-specific assignment for individual glycan structures attached a given amino acid sequence position.

3.10 Using the Glycoprotein Navigation Page to Find Global and Site-Specific Content

The glycoprotein summary page is a quick navigation aid to find information on global and site-specific glycosylation for individual and mixtures of glycoproteins. To access this information:

1. Point your browser at the UniCarbKB glycoprotein page at <http://www.unicarbkb.org/proteins> or select the “Glycoproteins” tab from the navigation bar.
2. To find a glycoprotein(s) of interest start typing the name(s) into the filter box, and UniCarbKB will match entries in the database (*see Note 9*). The updated table lists the UniProtKB accession number, taxonomy, reported number of glycan structures and a label indicating if site-specific glycosylation information is available.
3. To view information available for the glycoprotein(s) of interest, select the item(s) from the dropdown list, and press the Filter button.
4. The results table will list the matching UniCarbKB entries only. To access an individual glycoprotein summary page simple, click the name or accession number.

3.10.1 Glycoprotein Information

Every glycoprotein entry in UniCarbKB is linked to a comprehensive summary page, which makes it quick and easy for users to assess the known glycosylation status of a given protein. These pages provide a description of the attached glycan structures and knowledge of site-specific glycosylation that has been curated from the literature. When possible known site-specific data are mapped with data available in UniProtKB/Swiss-Prot. Figure 4 shows the glycoprotein summary page for human Alpha-2-HS-glycoprotein, at <http://www.unicarbkb.org/proteinsummary/P02765/annotated>. In summary, glycoprotein pages are split into two sections:

1. The main section displays information that is relevant to the glycoprotein:
 - The page starts with a header area with information about the protein, including its name, UniProtKB accession number linked to the UniProt website, and functional as well as PTM-related annotation imported from the UniProtKB/Swiss-Prot entry.
 - A summary of experimentally verified site-specific Glycosylation Sites clearly indicates the amino acid position and number of structures associated with that site.
 - Glycan structures reported at a specific site can be viewed by clicking the “Associated Structures” label.
 - An optional section shows compositional data if available.
 - All the glycan structures associated with the specified glycoprotein are listed below the Glycosylation Sites subsection.

(a) **Associated Structures**
Accession: P02765
UNIPROT/SWISS-PROT ENTRY

UniProtKB/Swiss-Prot PTM Description
Promotes endocytosis, possesses opsonic properties and influences the mineral phase of bone. Shows affinity for calcium and barium ions
Phosphorylated by FAM20C in the extracellular medium O- and N-glycosylated. O-glycosylated with core 1 or possibly core 8 glycans. N-glycan at Asn-156: Hex5HexNAc4; N-glycan heterogeneity at Asn-176: Hex5HexNAc4 (major) and Hex6HexNAc5 (minor)

(b) **Glycosylation Sites**

Position	Structures	Description	Evidence
ASN-156, ASN-176, THR-256, THR-270, SER-346	ASSOCIATED STRUCTURES 2	ELONG	GlycoSuite

Site-Specific Information
A number of glycan structures have been assigned to specific glycosylation sites

Position	Structures	Description	Evidence
ASN-156 AND ASN-176	ASSOCIATED STRUCTURES 1	SITE SPECIFIC	GlycoSuite
THR-256, THR-270 AND SER-346	ASSOCIATED STRUCTURES 1	SITE SPECIFIC	GlycoSuite

(c) **Compositional data available: 8 compositions reported:**

- 176 [Hex]5 [HexNAc]4 [Hex]5 [HexNAc]4 [NeuAc]2
- 346 [Hex]1 [HexNAc]1
- 156 [Hex]5 [HexNAc]4 [NeuAc]2 [Hex]4 [HexNAc]3 [NeuAc]1 [Hex]6 [HexNAc]5 [NeuAc]3 [Hex]6 [HexNAc]5 [Hex]1 [NeuAc]3 [Hex]5 [HexNAc]4

Glycan Structures

(d) **Biological Associations**
TAXONOMY (0) | PROTEIN (0) | SOURCE (0)

References 4

1. Comparison of sialylated N-glycopeptide levels in serum of pancreatic cancer patients, acute pancreatitis patients, and healthy controls
Kotro H, Joensuu S, Haglund C, Renkonen R
PubMed: 24841998 Year: 2014
2. Structure of the N- and O-glycans of the A-chain of human plasma alpha 2HS-glycoprotein as deduced from the chemical compositions of the derivatives prepared by stepwise degradation with exoglycosidases.
Watzlawick H, Walsh M, Yoshioka Y, Schmid K, Brossmer R
PubMed: 1457416 Year: 1992
3. Human urinary glycoproteomics; attachment site specific analysis of N- and O-linked glycosylations by CID and ECD
Halm A, Nilsson J, Ruetzsch U, Hesse C, Larson G
PubMed: 22171300 Year: 2012

Fig. 4 Screenshot for the human Alpha-2-HS-glycoprotein (UniProtKB/SwissProt:P02765). (a) A description of the glycoprotein is obtained from the UniProtKB Function and PTM/Processing sections in addition to the protein primary sequence. To obtain this information, we use the UniProtJAPI client [38]. (b) The “Associated Structures” label(s) provides information on level of global and/or site-specific information available and corresponding structures at each site for the Alpha-2-HS-glycoprotein entry. The level of information displayed can vary; for example, not all glycoproteins have site-specific data, however, all protein summary pages list (c) the glycan structures or compositions characterized at the global level, i.e. determined following an enzymatic or chemical release of glycans from a single purified glycoprotein. Finally, details on the relevant biological associations can be displayed by clicking the (d) “Taxonomy”, “Protein”, or “Source” boxes. All information presented on the protein summary pages have been sourced from the publications listed in the right-hand sidebar

For more information on the glycan structures, including supporting publications and biological associations, users are prompted to click the structure of interest.

2. The sidebar shows four sections:

- The “Structure Format” sidebar panel described in Subheading 3.6.1 to convert between the “CFG/Essentials”, “Oxford” and “Traditional” formats.
- The primary amino acid sequence of the protein, which is dynamically loaded from UniProtKB.
- Information regarding “Biological Associations” including taxonomy and tissue source, which provides a comprehensive summary of associated metadata.

- Peer-reviewed publications that were curated to compile the global and site-specific glycosylation data. For further details on the publication cited, users can navigate to the Reference page by clicking the publications title.

For example, Fig. 4 shows the summary page for alpha-2-HS-glycoprotein, which provides a description of the attached glycan structures and knowledge of site-specific glycosylation that has been curated from the literature. In addition, a comprehensive summary of associated metadata including biological source and publications citing the data is shown. Additional information about the glycoprotein can be displayed in UniProtKB/Swiss-Prot by clicking the listed protein accession identifier, P02765. The PTM/Processing section of this Swiss-Prot entry (http://www.uniprot.org/uniprot/P02765#ptm_processing) contains annotations for a number of glycosylation sites. Most of these sites have experimental evidence, and some of them are directly linked to site-specific pages in UniCarbKB via the “CAR_XXXXX” identifiers in the “Feature identifier” column, e.g. “Tyr-176 N-linked (GlcNAc...) (complex) CAR_000065”, which links to http://unicarbk.org/swissprotFT?feature=CAR_000065. The PTM/Processing section in Swiss-Prot also includes the global link to UniCarbKB entry (under “PTM databases”), as well as the free text comment about phosphorylation and glycosylation, which in turn is reproduced at the top of the UniCarbKB glycoprotein page (<http://unicarbk.org/proteinsummary/P02765/annotated>), as mentioned above.

3.11 Cell Lines

There are numerous challenges facing UniCarbKB’s growth and goal to organize and annotate mammalian glycoproteins. This is exemplified by the volume and diversity of data generated from cell line analysis, which has motivated us to create a dedicated cell line section (<http://www.unicarbk.org/celllines>) to display data on the relative abundance of glycan structures characterized from a small but growing number of cell line studies [39]. To date, the library is focused on colorectal cancer that includes over 25 primary cell lines representative of moderately differentiated metastatic, moderately differentiated primary, and poorly differentiated (aggressive) tumors [40]. Additional cell lines include a comparison between prostate cancer models from the work of Shah et al. [41].

4 Notes

1. Information presented in this chapter is based on the UniCarbKB March 2016 release. Regular updates and improvements are made to the user-interfaces and database model; therefore, web page content and design may have changed since publication. As part of our effort to provide

users with accurate information, updates to the interface and database content are described on the project user knowledge-base (<http://confluence.unicarbk.org>).

2. Results are displayed in a context-dependent manner. Some results will have descriptive text details, whereas others may link directly to the glycan structure or glycoprotein entry. In all cases, the search terms will be highlighted and appear in the title or supporting text of each result page.
3. As you type in the search box, you can find information quickly by seeing terms and keywords that match your search criteria. For example, as shown in Fig. 1, it is possible to search by multiple glycoproteins and when you start to type, “major” protein names partially matching the entered text will be listed. This feature serves two major advantages notably:
 - Save time searching—Choose from keywords to find information faster while typing less.
 - Spelling corrections—When searching for keyword, auto-complete will show matching or similar terms, this is particularly useful in the case of abbreviated classifications.
4. If searching for viruses, the species field describes the organism from which the DNA of the protein originates.
5. This option will only appear when the size of the structure dataset exceeds ten.
6. When searching by composition it is essential to choose *N*-linked or *O*-linked, otherwise no results will be shown.
7. Only those eukaryotic monosaccharides and modifications with supporting structure entries are listed, for example, glucosamine (GlcN) is not included since bacterial lipopolysaccharides are not stored in UniCarbKB.
8. The links in the sidebar change depending on the section of database you are viewing. This area provides important links to internal and external pages.
9. UniCarbKB also supports searching by UniProt accession number. To access this feature, use the “Protein Accession” option listed under the “Biological Source” menu, and follow the steps described in Subheading 3.10.1 substituting glycoprotein name with the UniProt accession identifier.

5 Summary

We are now at a critical point in the development of glyco-related databases. Continuing advancements in analytical technologies and new data types are unraveling the complexity of the glycome and glycoproteome; therefore, it is increasingly necessary to ensure all the data are annotated in a correct, consistent, and sufficient way.

UniCarbKB is an important step in this direction, with a continually expanding and carefully reviewed collection of glycan structures, annotated mammalian glycoproteins, and biological contextual information. As a valuable data storage and search platform, UniCarbKB is ready to embrace advances in user-interfaces and computational tools to best exploit accumulating glycoscience data and continue to enrich our understanding of glycosylation.

5.1 Submission of Updates, New Data, and Troubleshooting

The initiative is driven as a community endeavor and the team encourages end-user feedback. To submit updates and/or corrections to UniCarbKB and for any enquiries use the e-mail address info@unicarbkb.org. For more information and guides, please refer to our documentation site (<http://confluence.unicarbkb.org>).

Funding

UniCarbKB acknowledges funding from National eResearch Collaboration Tools and Resources project (Nectar) supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS). M.P.C. and N.H.P. acknowledge funding from a Macquarie University-Ludger Ltd Enterprise Partnership Scheme. The Swiss-Prot and Proteome Informatics groups are part of the SIB Swiss Institute of Bioinformatics supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI. The Swiss-Prot group also belongs to the UniProt Consortium funded by the National Institutes of Health (NIH) grant U41HG007822.

References

1. Dwek RA (1996) Glycobiology: toward understanding the function of sugars. *Chem Rev* 96(2):683–720
2. Haltiwanger RS, Lowe JB (2004) Role of glycosylation in development. *Annu Rev Biochem* 73:491–537. doi:[10.1146/annurev.biochem.73.011303.074043](https://doi.org/10.1146/annurev.biochem.73.011303.074043)
3. Ohtsubo K, Marth J (2006) Glycosylation in cellular mechanisms of health and disease. *Cell* 126(5):855–867
4. Campbell M, Hayes C, Struwe W, Wilkins M, Aoki-Kinoshita K, Harvey D, Rudd P, Kolarich D, Lisacek F, Karlsson N, Packer N (2011) UniCarbKB: putting the pieces together for glycomics research. *Proteomics* 11(21):4117–4121
5. Campbell MP, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita KF, Lisacek F, Packer NH (2014) UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res* 42(1):D215–D221. doi:[10.1093/nar/gkt1128](https://doi.org/10.1093/nar/gkt1128)
6. von der Lieth CW, Freire AA, Blank D, Campbell MP, Ceroni A, Damerell DR, Dell A, Dwek RA, Ernst B, Fogh R, Frank M, Geyer H, Geyer R, Harrison MJ, Henrick K, Herget S, Hull WE, Ionides J, Joshi HJ, Kamerling JP, Leeftang BR, Lutteke T, Lundborg M, Maass K, Merry A, Ranzinger R, Rosen J, Royle L, Rudd PM, Schloissnig S, Stenutz R, Vranken WF, Widmalm G, Haslam SM (2011) EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology* 21(4):493–502. doi:[10.1093/glycob/cwq188](https://doi.org/10.1093/glycob/cwq188)
7. Cooper C, Harrison M, Wilkins M, Packer N (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res* 29(1):332–335

8. UniProt C (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43(Database issue):D204–D212. doi:[10.1093/nar/gku989](https://doi.org/10.1093/nar/gku989)
9. York WS, Agravat S, Aoki-Kinoshita KF, McBride R, Campbell MP, Costello CE, Dell A, Feizi T, Haslam SM, Karlsson N, Khoo KH, Kolarich D, Liu Y, Novotny M, Packer NH, Paulson JC, Rapp E, Ranzinger R, Rudd PM, Smith DF, Struwe WB, Tiemeyer M, Wells L, Zaia J, Kettner C (2014) MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology* 24(5):402–406. doi:[10.1093/glycob/cwu018](https://doi.org/10.1093/glycob/cwu018)
10. Ranzinger R, Aoki-Kinoshita KF, Campbell MP, Kawano S, Lutteke T, Okuda S, Shinmachi D, Shikanai T, Sawaki H, Toukach P, Matsubara M, Yamada I, Narimatsu H (2015) GlycoRDF: an ontology to standardize glycomics data in RDF. *Bioinformatics* 31(6):919–925. doi:[10.1093/bioinformatics/btu732](https://doi.org/10.1093/bioinformatics/btu732)
11. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) PubChem substance and compound databases. *Nucleic Acids Res* 44(D1):D1202–D1213. doi:[10.1093/nar/gkv951](https://doi.org/10.1093/nar/gkv951)
12. Mariethoz J, Khatib K, Alloci D, Campbell MP, Karlsson NG, Packer NH, Mullen EH, Lisacek F (2016) SugarBindDB, a resource of glycan-mediated host-pathogen interactions. *Nucleic Acids Res* 44(D1):D1243–D1250. doi:[10.1093/nar/gkv1247](https://doi.org/10.1093/nar/gkv1247)
13. Shakhsheer B, Anderson M, Khatib K, Tadoori L, Joshi L, Lisacek F, Hirschman L, Mullen E (2013) SugarBind database (SugarBindDB): a resource of pathogen lectins and corresponding glycan targets. *J Mol Recognit* 26(9):426–431. doi:[10.1002/jmr.2285](https://doi.org/10.1002/jmr.2285)
14. Hayes C, Karlsson N, Struwe W, Lisacek F, Rudd P, Packer N, Campbell M (2011) UniCarb-DB: a database resource for glycomic discovery. *Bioinformatics* 27(9):1343–1344
15. Struwe WB, Pagel K, Benesch JL, Harvey DJ, Campbell MP (2015) GlycoMob: an ion mobility-mass spectrometry collision cross section database for glycomics. *Glycoconj J*. doi:[10.1007/s10719-015-9613-7](https://doi.org/10.1007/s10719-015-9613-7)
16. Aoki-Kinoshita K, Agravat S, Aoki NP, Arpinar S, Cummings RD, Fujita A, Fujita N, Hart GM, Haslam SM, Kawasaki T, Matsubara M, Moreman KW, Okuda S, Pierce M, Ranzinger R, Shikanai T, Shinmachi D, Solovieva E, Suzuki Y, Tsuchiya S, Yamada I, York WS, Zaia J, Narimatsu H (2016) GlyTouCan 1.0—the international glycan structure repository. *Nucleic Acids Res* 44(D1):D1237–D1242. doi:[10.1093/nar/gkv1041](https://doi.org/10.1093/nar/gkv1041)
17. Kirschner KN, Yongye AB, Tschampel SM, Gonzalez-Outeirino J, Daniels CR, Foley BL, Woods RJ (2008) GLYCAM06: a generalizable biomolecular force field. *Carbohydrates*. *J Comput Chem* 29(4):622–655. doi:[10.1002/jcc.20820](https://doi.org/10.1002/jcc.20820)
18. Alloci D, Mariethoz J, Horlacher O, Bolleman JT, Campbell MP, Lisacek F (2015) Property graph vs RDF triple store: a comparison on glycan substructure search. *PLoS One* 10(12):e0144578. doi:[10.1371/journal.pone.0144578](https://doi.org/10.1371/journal.pone.0144578)
19. Cooper CA, Gasteiger E, Packer NH (2001) GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* 1(2):340–349. doi:[10.1002/1615-9861\(200102\)1:2<340::AID-PROT340>3.0.CO;2-B](https://doi.org/10.1002/1615-9861(200102)1:2<340::AID-PROT340>3.0.CO;2-B)
20. Agravat SB, Saltz JH, Cummings RD, Smith DF (2014) GlycoPattern: a web platform for glycan array mining. *Bioinformatics* 30(23):3417–3418. doi:[10.1093/bioinformatics/btu559](https://doi.org/10.1093/bioinformatics/btu559)
21. Gotz L, Abrahams JL, Mariethoz J, Rudd PM, Karlsson NG, Packer NH, Campbell MP, Lisacek F (2014) GlycoDigest: a tool for the targeted use of exoglycosidase digestions in glycan structure determination. *Bioinformatics* 30(21):3131–3133. doi:[10.1093/bioinformatics/btu425](https://doi.org/10.1093/bioinformatics/btu425)
22. Campbell M, Royle L, Radcliffe C, Dwek R, Rudd P (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* 24(9):1214–1216
23. Ranzinger R, Herget S, Wetter T, von der Lieth C (2008) GlycomeDB—integration of open-access carbohydrate structure databases. *BMC Bioinformatics* 9:384
24. Toukach PV, Egorova KS (2016) Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. *Nucleic Acids Res* 44(D1):D1229–D1236. doi:[10.1093/nar/gkv840](https://doi.org/10.1093/nar/gkv840)
25. Harvey DJ, Merry AH, Royle L, Campbell MP, Dwek RA, Rudd PM (2009) Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds. *Proteomics* 9(15):3796–3801. doi:[10.1002/pmic.200900096](https://doi.org/10.1002/pmic.200900096)
26. Harvey DJ, Merry AH, Royle L, Campbell MP, Rudd PM (2011) Symbol nomenclature for representing glycan structures: extension to cover different carbohydrate types. *Proteomics* 11(22):4291–4295. doi:[10.1002/pmic.201100300](https://doi.org/10.1002/pmic.201100300)
27. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R (2006) Advancing glycomics: implementation strategies at the Consortium for Functional

- Glycomics. *Glycobiology* 16(5):82R–90R. doi:[10.1093/glycob/cwj080](https://doi.org/10.1093/glycob/cwj080)
28. Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME (2009) Symbol nomenclature for glycan representation. *Proteomics* 9(24):5398–5399. doi:[10.1002/pmic.200900708](https://doi.org/10.1002/pmic.200900708)
 29. Varki A, Cummings RD, Aebi M, Packer NH, Seeberger PH, Esko JD, Stanley P, Hart G, Darvill A, Kinoshita T, Prestegard JJ, Schnaar RL, Freeze HH, Marth JD, Bertozzi CR, Etzler ME, Frank M, Vliegenthart JF, Lutteke T, Perez S, Bolton E, Rudd P, Paulson J, Kanehisa M, Toukach P, Aoki-Kinoshita KF, Dell A, Narimatsu H, York W, Taniguchi N, Kornfeld S (2015) Symbol nomenclature for graphical representations of glycans. *Glycobiology* 25(12):1323–1324. doi:[10.1093/glycob/cwv091](https://doi.org/10.1093/glycob/cwv091)
 30. Herget S, Ranzinger R, Maass K, von der Lieth C (2008) GlycoCT—a unifying sequence format for carbohydrates. *Carbohydr Res* 343(12):2162–2171
 31. Banin E, Neuberger Y, Altshuler Y, Halevi A, Inbar O, Dotan N, Dukler A (2002) A novel LinearCode(R) nomenclature for complex carbohydrates. *Trends Glycosci Glycotechnol* 14(77):127–137
 32. Aoki K, Yamaguchi A, Ueda N, Akutsu T, Mamitsuka H, Goto S, Kanehisa M (2004) KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res* 32(Web Server issue):W267–W272
 33. Bohne-Lang A, Lang E, Forster T, von der Lieth C (2001) LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res* 336(1):1–11
 34. Tanaka K, Aoki-Kinoshita KF, Kotera M, Sawaki H, Tsuchiya S, Fujita N, Shikanai T, Kato M, Kawano S, Yamada I, Narimatsu H (2014) WURCS: the Web3 unique representation of carbohydrate structures. *J Chem Inf Model* 54(6):1558–1566. doi:[10.1021/ci400571e](https://doi.org/10.1021/ci400571e)
 35. Campbell M, Ranzinger R, Lutteke T, Mariethoz J, Hayes C, Zhang J, Akune Y, Aoki-Kinoshita K, Damerell D, Carta G, York W, Haslam S, Narimatsu H, Rudd P, Karlsson N, Packer N, Lisacek F (2014) Toolboxes for a standardised and systematic study of glycans. *BMC Bioinformatics* 15(Suppl 1):S9
 36. Ceroni A, Dell A, Haslam S (2007) The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* 2:3
 37. Damerell D, Ceroni A, Maass K, Ranzinger R, Dell A, Haslam SM (2012) The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments. *Biol Chem* 393(11):1357–1362. doi:[10.1515/hsz-2012-0135](https://doi.org/10.1515/hsz-2012-0135)
 38. Patient S, Wieser D, Kleen M, Kretschmann E, Jesus Martin M, Apweiler R (2008) UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics* 24(10):1321–1322. doi:[10.1093/bioinformatics/btn122](https://doi.org/10.1093/bioinformatics/btn122)
 39. Campbell MP, Packer NH (2016) UniCarbKB: new database features for integrating glycan structure abundance, compositional glycoproteomics data, and disease associations. *Biochim Biophys Acta*. doi:[10.1016/j.bbagen.2016.02.016](https://doi.org/10.1016/j.bbagen.2016.02.016)
 40. Holst S, Deuss AJ, van Pelt GW, van Vliet SJ, Garcia-Vallejo JJ, Koeleman CA, Deelder AM, Mesker WE, Tollenaar RA, Rombouts Y, Wuhrer M (2016) N-glycosylation profiling of colorectal cancer cell lines reveals association of fucosylation with differentiation and Caudal Type Homeobox 1 (CDX1)/Villin mRNA expression. *Mol Cell Proteomics* 15(1):124–140. doi:[10.1074/mcp.M115.051235](https://doi.org/10.1074/mcp.M115.051235)
 41. Shah P, Wang X, Yang W, Toghi Eshghi S, Sun S, Hoti N, Chen L, Yang S, Pasay J, Rubin A, Zhang H (2015) Integrated proteomic and glycoproteomic analyses of prostate cancer cells reveal glycoprotein alteration in protein abundance and glycosylation. *Mol Cell Proteomics* 14(10):2753–2763. doi:[10.1074/mcp.M115.047928](https://doi.org/10.1074/mcp.M115.047928)

Impact of Nonsynonymous Single-Nucleotide Variations on Post-Translational Modification Sites in Human Proteins

Naila Gulzar, Hayley Dingerdissen, Cheng Yan, and Raja Mazumder

Abstract

Post-translational modifications (PTMs) are covalent modifications that proteins might undergo following or sometimes during the process of translation. Together with gene diversity, PTMs contribute to the overall variety of possible protein function for a given organism. Single-nucleotide polymorphisms (SNPs) are the most common form of variations found in the human genome, and have been found to be associated with diseases like Alzheimer's disease (AD) and Parkinson's disease (PD), among many others. Studies have also shown that non-synonymous single-nucleotide variation (nsSNV) at the PTM site, which alters the corresponding encoded amino acid in the translated protein sequence, can lead to abnormal activity of a protein and can contribute to a disease phenotype. Significant advances in next-generation sequencing (NGS) technologies and high-throughput proteomics have resulted in the generation of a huge amount of data for both SNPs and PTMs. However, these data are unsystematically distributed across a number of diverse databases. Thus, there is a need for efforts toward data standardization and validation of bioinformatics algorithms that can fully leverage SNP and PTM information for biomedical research. In this book chapter, we will present some of the commonly used databases for both SNVs and PTMs and describe a broad approach that can be applied to many scenarios for studying the impact of nsSNVs on PTM sites of human proteins.

Key words Disease association, Pan-cancer variomes, Post-translational modifications (PTMs), PTM databases, Single-nucleotide variations, SNP databases, nsSNVs

1 Introduction

The discovery of the helical structure of DNA in 1953 by James Watson and Francis Crick was critical to promoting a better understanding of the complexities of molecular biology. Shortly thereafter, Crick introduced the “central dogma” describing the flow of information in a biological system from DNA to RNA to protein. He later wrote [1]: *“The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.”*

Almost 60 years later, the central dogma continues to dictate the majority of biological information transfer. However, there are

now well-supported exceptions to this model that implicate protein-mediated information transfer [2]. Research on bovine spongiform encephalopathy (BSE), also known as mad cow disease, and the related human Creutzfeldt-Jakob disease (CJD) have implicated a proteinaceous infectious particle, or prion, in disease pathogenesis [3]. Although not all prions have a negative impact, the BSE and CJD prions can exist in two conformations: a soluble, ubiquitous, and seemingly innocuous protein, and an insoluble form that promotes aggregation of the prions [4]. Once aggregated, these proteins can induce their soluble counterparts to adopt the pathogenic aggregate form in turn [5].

Thus, prions present a clear violation of the central dogma. In fact, it is thought that this may be one of the most extreme among a number of mechanisms by which information transfer can occur, the spectrum of epigenetic variation presenting additional challenges to the theory of the central dogma [2]. However, while this dogma is not an absolute principle, it does hold true for the majority of known information transfer. The protein modifications and subsequent impact thereof discussed throughout this chapter assume a central dogma model: alteration of the nucleotide sequence results in altered flow of information to the protein, which may, in turn, affect any post-translational modifications (PTMs) normally occurring on that protein.

1.1 Genetic Variations, Mutations, and Single-Nucleotide Polymorphisms (SNPs)

Genetic variation includes Sequence ontology identifier SO:0000694 (SNPs) and structural variants that are heritable, as well as mutations that may or may not be heritable. SNPs are the most common form of variation, distributed throughout the entire genome with an average density of one SNP per 1.9 kilobases [6, 7]. According to the dbSNP database [8], there are currently over 100 million SNPs documented in the human genome. A major challenge to the comprehensive cataloging of SNPs is the difficulty in detecting very rarely occurring SNPs. Although the 1000 Genomes Project has identified some polymorphisms whose minor allele frequency (MAF) is less than 5 % [9], it is expected that there are actually millions of human SNPs with an even lower MAF that may not yet be accounted for in SNP databases [7]. SNPs are not necessarily disease-associated, but many disease associations have been linked to SNPs [10, 11].

While SNPs refer to variations at the population level, somatic mutations are changes that occur in an individual over the course of a lifetime. Somatic mutations do not have to be deleterious, but they certainly can be harmful, as in the case of recurrent mutations in the p53 (TP53) gene associated with cancer [12]. Mutations can also occur in germ cells, called germline mutations, hence they can be passed to their progeny. A germline mutation is a rare event: when a variant is found in a population with a very low frequency, for example, less than 1 %, this may be an indication that the variant is a result of recent genetic changes in the germline of that

population. A variation with MAF greater than 5 % is considered a common variant [13]. Throughout this chapter, we will use SNP in reference to well-characterized variants in a population, and we will use the term single-nucleotide variation (SNV SO:0001483) to describe rarer variants that have been observed in one or several individuals but have not yet been well characterized in any population.

1.2 Classification of Mutations

When a single nucleotide base in the DNA sequence is altered, it is termed a **point mutation** (Sequence Ontology accession SO:1,000,008 [14]). Such mutations may occur due to (1) intrinsic factors, such as damage during replication that escaped the DNA repair machinery or interaction with free radicals generated during metabolism, or (2) extrinsic factors, like exposure to UV radiation or chemical mutagens. Point mutations can be classified by their effects? on sequence structure, chemical composition, impact on function, and cellular location in which they occur. Our focus is primarily on the classification of mutations occurring in coding regions and their impact on the translated products of these regions, but it is important to keep in mind that such mutations can also occur in non-coding regions of DNA. Mutations occurring outside of coding regions are expected to be silent unless they affect translation machinery binding, resulting in potential dysregulation of protein synthesis, or they alter splice sites, potentially resulting in novel transcripts. Refer to Fig. 1 for the classification of point mutations.

1.2.1 Structural Consequences

Point mutations involving the exchange of one nucleotide for another are also called single base substitution mutations. Because the number of nucleotides in the DNA sequence remains the same for this kind of mutation, the translation frame remains intact. Insertions and deletions (indels) of a single nucleotide are still considered point mutations, but due to the resulting increase or decrease in the number of nucleotides in the sequence, the translation frame may be shifted, usually resulting in a change in the corresponding codons. Thus, these mutations belong to the larger class of so-called frameshift mutations.

1.2.2 Chemical Composition

Nucleotide substitutions are further characterized with respect to the chemistry of the specific nucleotides involved.

- (a) **Transition (SO Accession: SO:1,000,009)**: These mutations result when a purine nucleotide base is replaced by another purine, or when a pyrimidine is substituted by another pyrimidine. Adenine replaced by guanine or cytosine by thymine are transition mutations.
- (b) **Transversion (SO Accession: SO:1,000,017)**: These mutations result when a purine is replaced by a pyrimidine in the DNA, or a pyrimidine by a purine, for example, when adenine is replaced by cytosine or guanine by thymine.

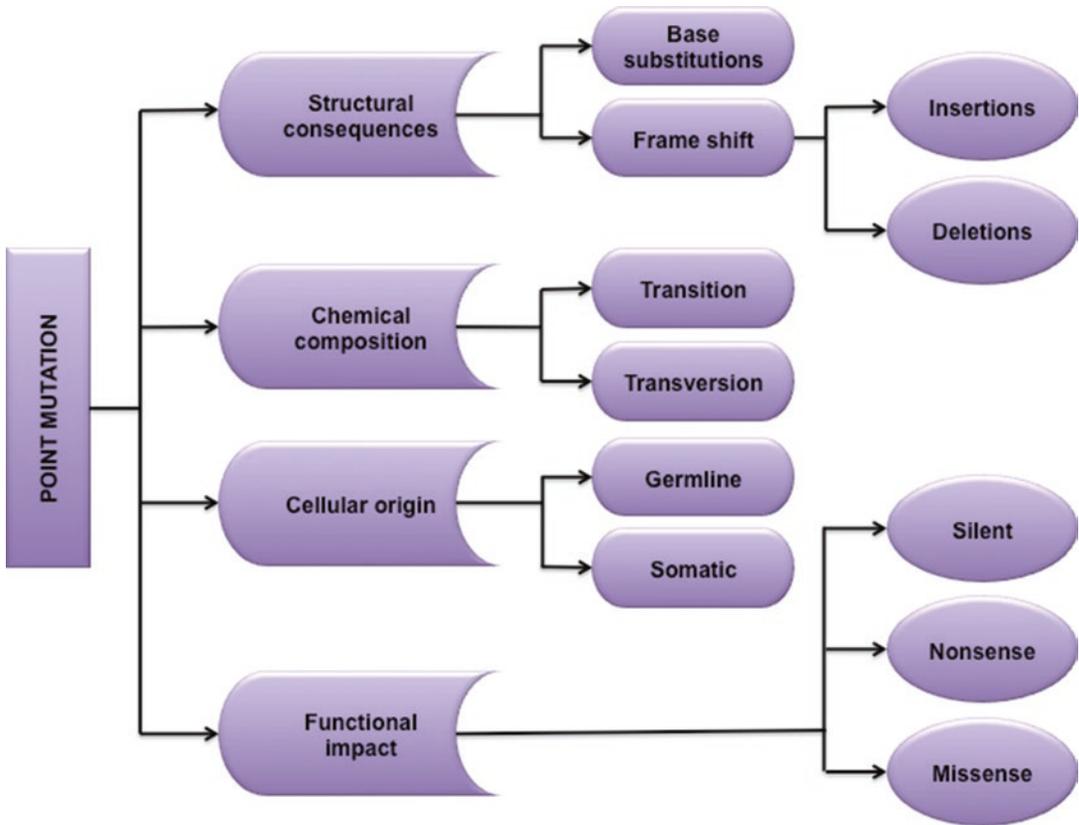


Fig. 1 Classification of point mutations

Although the ratio of transitions to transversions (T_s/T_v) is expected to be 0.5 if we assume no biological bias exists, the observed T_s/T_v ratio is generally greater than 0.5: this is due to tautomeric shifts and the fact that transitions more frequently result in tolerable silent mutations (SO Accession: SO:0,001,017) as explained by the wobble hypothesis [15].

1.2.3 Functional Impact

- (a) **Silent mutation:** This type of base substitution in coding regions does not lead to any change in the resulting amino acid of the translated protein sequence due to the degeneracy of the genetic code. The genetic code is written in nucleotide triplets: as there are only four nucleotides, there are $4^3 = 64$ possible codons to encode the 20 amino acids. In fact, 61 of the triplet combinations encode the amino acids while the remaining three encode stop codons, which mark the termination of translation. This concept of degeneracy, especially with respect to flexibility in the third position of a codon, is otherwise known as the wobble hypothesis [16], and lowers the risk of deleterious effects caused by point mutations by increasing the likelihood that a mutation will have a silent phenotype. Silent mutations are often referred to as synonymous mutations (SO

Accession: SO:0,001,819) when the encoded residue is the same in the variant product as in the normal product, despite two distinct but synonymous codons. Silent mutations, therefore, also result in the preservation of protein function.

- (b) **Nonsense mutation (SO Accession: SO:1,000,062)**: Those base substitutions that result in a premature stop codon in the coding region of the nucleotide sequence are termed nonsense mutations. This type of mutation results in truncated proteins, which generally causes a loss of function. A well-studied case involves the group of thalassemias which are disorders caused by heritable mutations including deletions of key fragments in the α - and β -globin genes. These mutations result in the formation of abnormal hemoglobin, which manifests as anemia [17].
- (c) **Missense mutation (SO Accession: SO:0,001,583)**: This mutation occurs when one nucleotide is replaced by another in such a way that the new codon formed results in the incorporation of a different amino acid in the translated protein but the length of the protein is preserved. Missense mutations can be considered **conservative** (SO:1,000,060) when the replacement of the nucleotide results in the replacement of amino acid with similar chemical properties because its function is less likely to be altered. For example, a polar amino acid replaced by another polar amino acid in the protein sequence is considered to be a conservative missense mutation. A **non-conservative** or non-synonymous missense mutation (SO:1,000,061), however, results in the translation of an amino acid with different chemical properties, and is therefore likely to alter protein function. A polar amino acid replaced by a hydrophobic amino acid is an example of a non-conservative missense mutation (SO:0,001,586). Although both nonsense and missense mutations can be considered non-synonymous, for the duration of this paper, we refer specifically to missense variants as non-synonymous single nucleotide variations, or nsSNVs.

1.2.4 Cellular Origin

As described above, germline and somatic mutations are characterized based on whether the mutation occurs and is propagated to the germ cells or non-reproductive cells of the body, respectively. Since gametes develop from germ cells, this type of mutation can be passed onto progeny. However, since somatic cells do not give rise to gametes, these mutations are not passed onto progeny. Cancer-causing somatic mutations commonly arise in proto-oncogenes, which are frequently involved in regulation of cell division and, when mutated, lead to uncontrolled cell division and formation of tumor [18].

1.3 Co- and Post-translational Modifications (PTMs)

Proteins undergo many types of covalent modifications, both during and after the process of translation, which result in the diversification of protein function. In conjunction with the diversity of genes

present in an organism, covalent modifications catalyzed by specific enzymes are responsible for the biological complexity and regulatory functions characteristic of advanced species. Co- and post-translational modifications (PTMs) can occur on the amino- (N-) or carboxy- (C-) terminus of the polypeptide, or on the side chain of the protein. In addition to diversification, PTMs result in distinct roles in processes, like cellular differentiation and development, protein degradation, and cell adhesion [19]. For instance, histone proteins, which are integral to chromatin formation, undergo many PTMs in the N-terminus, thereby regulating access to bound DNA and inducing specific transcription “on-off” states. Histone modifications together with epigenetic mechanisms, like DNA methylation, contribute to an additional level of complexity of the genome and augment the total information potential of the genetic code [20, 21].

Depending on the functional group attached to the protein, PTMs are classified into different groups. Attachment of specific functional groups, like phosphate, sugar, acetate, amide, or methyl groups, affects the chemical, kinetic, and structural properties of the protein. The most common form of PTM is phosphorylation and the most complex in terms of structure and diversity is glycosylation [22]. It should be noted that the spectrum of PTM is much broader than presented herein: the scope of this discussion is limited to a subset of the most frequent PTMs as counted by occurrences in the UniProt/Swiss-Prot Knowledgebase [23].

1.3.1 Phosphorylation

Protein phosphorylation was first reported by Phoebus Levene in 1906, but it would take almost 50 years for the kinase mechanism of phosphorylation to be discovered [24]. This biochemical process is the most extensive type of PTM used in signal transduction [25, 26]. Phosphorylation is integral to all basic cellular processes, including signal transduction, metabolism, growth, division, differentiation, organelle trafficking, muscle contraction, immunity, learning, and memory [25, 27, 28]. Post-translational phosphorylation involves the transfer of a phosphate group from ATP to the side chains of proteins on various residues, resulting in the formation of the phosphoamino acids listed in Table 1. Although

Table 1
Phosphoamino acids resulting from attachment of phosphate group to amino acid

Class of phosphoamino acids	Amino acids involved
Acyl phosphates	Aspartate and glutamate
Phosphomonoesters	Serine, threonine, and tyrosine
Phosphoramidates	Histidine, arginine, and lysine
Thiophosphate	Cysteine

phosphorylation on a variety of residues is possible, serine (Ser), threonine (Thr), and tyrosine (Tyr)-based phosphorylations have dominated research as these modifications are acid stable and can be studied by using acid treatment methods [29]. By contrast, the phosphoramidates undergo hydrolysis under acidic conditions and therefore have significantly less representation in research [29].

Phosphorylation is catalyzed by the specific actions of two types of enzymes: kinases and phosphatases. Kinases are the enzymes responsible for phosphorylating the proteins, whereas phosphatases are the enzymes that cleave the phosphate group from phosphorylated proteins. Reversible phosphorylation of proteins on Ser, Thr, and Tyr results in a dynamic equilibrium between activated and deactivated states of the affected proteins through conformational changes induced by the presence or absence of the phosphate groups. For example, phosphorylation plays an important role in the regulation of signaling involving G-protein coupled receptors (GPCRs), such as by homologous desensitization [30]. In this type of regulation, receptor binding leads to the phosphorylation of residues at the C-terminal tail of GPCRs by their specific enzymes, G-protein-receptor kinases (GRKs). Phosphorylation of GRKs causes binding of β arrestins at the C-terminal tail of GPCRs leading to a conformational change in β arrestins and the binding of clathrin and other endocytic components [31]. Errors in the regulation of GPCRs have been linked to diseases like nephrogenic diabetes insipidus, retinitis pigmentosa, and morbid obesity [32–34]. Interestingly, studies have shown that the pattern of receptor phosphorylation generates a tissue-specific “bar code” to guide physiologically relevant receptor signaling [35]. GPCRs are part of a huge superfamily of cell surface signaling proteins, only a few of which have been targeted for drug discovery with good therapeutic outcome. Efforts are underway to target specific GPCRs for the prevention and treatment of cancer [36]. Some SNVs at phosphorylation sites may impact protein function and eventually lead to disease phenotype (*see Note 1*).

1.3.2 Glycosylation

Glycosylation is the process of modifying proteins or lipids by covalently attaching mono- or oligosaccharides to certain functional groups of amino acids. This modification plays a pivotal role in inter- and intracellular biological processes, such as signaling, protein maturation and turnover, cell adhesion and trafficking, receptor binding, and activation [37–41]. Glycans are created by non-template synthesis and provide cell- and tissue-specific information. Attachment of sugars generally occurs at specific residues on proteins, or to the growing end of the oligosaccharide chains. Glycosylation of proteins is a widespread and structurally diverse type of PTM [42]. Multiple factors responsible for the glycan biosynthetic potential of a given cell are: relative enzyme abundance and localization, abundance and trafficking of glycoprotein

substrates, and the localization and concentration of high-energy nucleotide sugar donors like UDP-*N*-acetylglucosamine [42]. The two main organelles of central importance to protein glycosylation are the endoplasmic reticulum (ER) and the Golgi apparatus. In the ER, glycosylation acts as a quality control mechanism because only properly folded proteins are transported to the Golgi. Sugar moieties are added to soluble proteins in the trans Golgi network to guide delivery to the proper destinations: these moieties can act as ligands for cell surface receptors to stimulate signal transduction pathways or cell attachment [41]. Although mammalian glycans can be conserved, there do exist species-specific variations that are thought to be involved in the emergence of distinct traits, like disease susceptibility, and can therefore be targeted for treatments [43–45]. For example, the clinical usefulness of glycans as blood antigens was proposed after the discovery of human blood groups [46], and the antithrombotic glycan heparin is one of the most widely used drugs because of its glycan-mediated anticoagulant activity [47]. New and improved methods of glycan analysis and advanced molecular modeling techniques are needed to better understand the full spectrum of glycoprotein functions and to identify glycosylation-based biomarkers of diseases. There are more than 270 human glycosyltransferases, the enzymes responsible for addition of the sugar group to the amino acid, and about 2 % of the human genome codes for proteins which are involved in glycan biosynthesis, degradation, or transport [48].

Depending on the nature of the glycosidic bond and the type of sugar group attached, there are many categories of glycoconjugates. The most studied are N- and O-linked glycoproteins, proteoglycans, GPI-anchored proteins, and glycolipids [49].

- (a) **N-linked glycosylation** is the most studied and common type of glycosylation [50], and is prominent in eukaryotic secretory proteins [51]. This PTM is initiated co-translationally on the luminal side of the ER while the protein is synthesized on the ribosomes attached to the ER. The enzyme oligosaccharyltransferase catalyzes the amide linkage between the 14-sugar glucose3-, mannose9-, *N*-acetylglucosamine2 (Glc3Man9GlcNAc2) core oligosaccharide, and the asparagine (Asn) residue of the glycosylation sequon, Asn-X-Ser/Thr (NXS/T). Here, X may represent any amino acid except proline (Pro), Ser, and Thr, because pro is believed to cause steric hindrance that impairs a protein's ability to stabilize the glycosidic bond formation [30, 51]. Presence of the glycosylation sequon does not guarantee glycosylation of the protein: some NXS/T sequons are not glycosylated because they are inaccessible to the responsible enzymes [52]. After multiple enzyme-catalyzed steps in the ER and processing in the Golgi, glycoproteins are transported to their destinations. Although glycans themselves are diverse, the dolichol-bound oligosaccha-

ride Glc3Man9GlcNAc2, acting as the precursor to initiate N-glycosylation in the ER, is conserved across Eukaryotes and Archaea [53]. The significance of N-linked glycosylation can be demonstrated again by consideration of GPCRs. One GPCR family member, protease-activated receptor-1 (PAR1), is activated by many specific proteases, including thrombin. The specificity of G-protein coupling and the differential regulation of cellular responses are associated with the N-linked glycosylation of PAR1. Thus, glycosylation of GPCRs specifies the coupling of distinct G-protein subtypes [47]. PAR1 is an important drug target for vascular and thrombotic diseases [54] and has been linked to the progression of some malignant cancers [55] (*see Note 2* for some valuable details on the possible effects of nsSNVs on N-linked glycosylation sites of the proteins).

- (b) **O-linked glycosylation** of Ser and Thr was discovered in the early 1980s. This type of glycosylation is typically found at the nuclear envelope and on nuclear proteins, transcription factors, and cytoskeletal proteins. Two enzyme complexes play a vital role in the regulation of O-linked glycosylation, namely O-linked N-acetylglucosamine transferase (OGT), which adds monosaccharide units to the hydroxyl group of the amino acids (generally Ser/Thr residues on the proteins), and O-GlcNAcase (OGA), which removes the O-GlcNAc moiety from proteins [49, 56]. In mammals, O-linked glycosylation involves diverse modifications and is generally classified based on the innermost or initiating monosaccharide. For example, in the mucin type of O-glycosylation, an α -linked N-acetylgalactosamine (GalNAc) is the initiating sugar moiety [56] (*see Note 3*).
- (c) **C-mannosylation** occurs with low frequency and involves a C–C bond between an α -mannosyl residue and the C-2 of tryptophan (Trp). Examples of this type of linkage are found in mammalian proteins, such as RNase2, interleukin-12, and properdin.

1.3.3 Ubiquitination

This PTM involves the conjugation of the polypeptide ubiquitin with the lysine (Lys) residue of the target protein using three different enzymes in tandem [57]. The first reaction is catalyzed by the ubiquitin-activation enzyme E1 that activates ubiquitin in an ATP-dependent process. The second enzyme is a ubiquitin conjugation enzyme or ubiquitin carrier enzyme (UBC or E2) that accepts ubiquitin from E1. The third enzyme is ubiquitin protein ligase, E3, which catalyzes the transfer of ubiquitin from E2 to the ϵ -amino group of the target protein lysine residue [58]. Ubiquitination is reversible through the action of deubiquitinases (DUBs), responsible for the removal of ubiquitin from the target proteins. The human genome encodes for two E1 enzymes, almost 38 E2 enzymes, about 600 to 1000 E3 enzymes [58], and almost

90 DUBs [58, 59]. Generally, this attachment of ubiquitin tags a protein for degradation by the 26S proteasome, but non-proteolytic roles have been found in the regulation of DNA repair, chromatin dynamics, protein kinase activation, membrane trafficking, and cell signaling [60]. The difference in the pattern of ubiquitination (mono- or polyubiquitination) and how the groups are attached to the receptor determines the fate of the target receptor [58]. The general mechanism of ubiquitination involves a ubiquitin or polyubiquitin signal to recruit proteins possessing ubiquitin-binding domains: ubiquitinated proteins and ubiquitinated receptors are then brought together to execute their specific functions [60, 61]. Notably, the deregulation of monoubiquitination of histone H2B protein has been associated with tumorigenesis. Normally, monoubiquitinated H2B functions as a transcriptional regulator and the enzymes which drive this modification are linked to cancer development [62] (*see Note 4* for statistics about ubiquitination sites in human proteins and the association of nsSNV-affected ubiquitination sites with disease susceptibility).

1.3.4 Acetylation

Acetylation involves the addition of an acetyl group to a target amino acid in a protein and plays a significant role in protein stability, localization, metabolism, and apoptosis (*see Note 5* for general statistical details about acetylation sites of human proteins). Acetylation occurs co- and post-translationally, and can be grouped into two types:

- (a) **N-terminal acetylation** (Nt-acetylation) is a co-translational modification occurring across all taxonomic kingdoms and is common in cytoplasmic eukaryotic proteins [63, 64]. Nt-acetylation is catalyzed by N-terminal acetyltransferases (NATs), which include a set of enzyme complexes that transfer an acetyl group from acetyl CoA (Ac-CoA) to the α -amino group of the first amino acid residue of the protein [64]. Ac-CoA has been shown to act as a signaling molecule and has been linked to the regulation of Nt-acetylation of apoptosis-regulating proteins [65], thus depicting an interesting role of Nt-acetylation in metabolic states and cell death. NAT function also appears to be important, based on the observation that a mutation in the NAT gene is responsible for the human hereditary lethal disease Ogden syndrome, which causes developmental delay and death in infancy [66]. As it turns out, acetylation is implicated in a diversity of regulatory functions, including regulation of protein degradation by regulating ligases [67], inhibition of protein translocation from the cytosol to the ER [68], and mediation of protein complex formation [69]. Nt-acetylation also mediates the attachment of small GTPases to membrane-associated proteins involved in organelle trafficking [70].

- (b) **Lysine acetylation** is the post-translational modification of proteins catalyzed by lysine acetyltransferase enzymes (KATs). Lysine acetylation was first discovered as a PTM of histones linked to the regulation of chromatin structure and function [71]. The specific enzymes involved in this modification are histone acetyltransferases (HATs) and histone deacetylases (HDACs). Subsequent research on HATs and HDACs showed that they play a role in transcription and other cellular processes like protein stability. Interestingly, in addition to the general HAT/HDAC functions, four HATs—CREB binding protein (CBP), p300 HAT, P300/CBP-associated factor (PCAF), and transcription initiation factor TFIID subunit 1 (TAF1)—and one HDAC, histone deacetylase 6 (HDAC6), have been found to possess ubiquitin-linked functions. These activities clearly depict the link between lysine acetylation and ubiquitination, which makes it an important element in the regulation of cellular proteolytic activities [71]. HATs and HDACs are named for their primary action on histones, but these enzymes have also been observed to modify acetyl groups on non-histone proteins. Studies have shown that HDAC inhibitors (HDACI) lead to the differential expression of proteins involved in various biological pathways, including cell cycle progression, apoptosis, free radical generation, autophagy, and DNA damage repair. A study on bladder cancer has demonstrated that HDACIs are associated with the inhibition of cell proliferation and induction of apoptosis in the bladder cancer cells by inactivation of HDACs or modulation of chromatin structure. This suggests the role of HDACIs in the development of therapeutic drugs for bladder cancer treatment [72, 73].

1.3.5 Methylation

This type of PTM occurs on all basic amino acid residues, and can be categorized into two subtypes depending on whether the modified residue is Lys, via lysine methyltransferases, or Arg, via protein arginine methyltransferases (PRMTs) [74–77]. Lys has been found to be monomethylated, meaning one methyl group (me1) is attached; dimethylated, with two methyl groups (me2) attached; or trimethylated, with three methyl groups (me3) attached to their ϵ -amine group. Arg can be monomethylated (me1), symmetrically dimethylated (me2a), or asymmetrically dimethylated (me2a) on their guanidinyll residue, and His has rarely been observed to be monomethylated. Nine mammalian PRMTs have been characterized to be involved in vital biological processes such as signal transduction, gene transcription, DNA repair, and mRNA splicing; studies have demonstrated their connection with carcinogenesis and metastasis [78]. Although methylation was initially considered to be an irreversible process [79], the discovery of histone H3 lysine 4 (H3K4) demethylase demonstrated that Lys methylation involves both methylases and demethylases [80] [81].

Among the most significant and extensively studied of known methylated proteins are the histones. Histone methylation sites at lysine residues include H3K4, H3K9, H2K27, H3K36, H3K79, and H3K20 [36], and histone arginine methylation sites include H3R2, H3R8, H3R17, H3R26, and H4R3. Many additional basic amino acid residues of histone proteins H1, H2A, H2B, H3, and H4 have also been found to be methylated as well [82]. As mentioned above, eukaryotic chromatin is packed with histones, which are subject to several types of PTMs, including methylation. These so-called epigenetic modifications extend the genetic information carried by the DNA sequence and constitute a “histone code”. Studies have shown that histone methylation sites on chromatin serve as the binding site for chromatin effector molecules and their associated complexes, thereby regulating transcription [36]. Thus, changes in histone methylation can have deleterious effects and have been implicated in human disease, particularly cancer.

Interestingly, methylation status can be maintained and passed onto cellular progeny via the inheritance of silenced heterochromatin through mitosis. In other instances, however, methylation patterns are dynamic and vary greatly at different stages of cell development, and can change in response to environmental signals during the cell differentiation process [81]. Missense mutations of Lys27Met (K27 M) and Gly34Arg/Val (G34R/V) in the genes that encode histone H3.3 (H3F3A) and H3.1 (HIST3H1B) have been observed in pediatric gliomas. This type of mutation has been found in almost 80 % of the cases of diffuse intrinsic pontine gliomas (DIPGs), incurable tumors of the brain stem in children [83] (*see Note 6* for additional possible effects of nsSNVs on methylation sites of human proteins and their associations with disease).

2 Materials

Having established the significance of nsSNVs and PTMs, we will now turn our attention to the various locations at which the relevant data can be found. Please see Table 2 for a summary of the various databases and the types of information they provide.

2.1 Variation Databases

There has been an exponential increase in the generation of sequencing data due to technological advances (e.g. next-generation sequencing (NGS)) over the past decade. Hundreds and thousands of variations have now been associated with diseases, revolutionizing the field of cancer biology and leading to the discovery of new biomarkers [84]. This huge amount of data is documented across several databases [85]. Some commonly used and well-trusted databases containing SNVs and related information include:

Table 2
Summary of variation and PTM databases

Variation databases	Database URLs
BioMuta	https://hive.biochemistry.gwu.edu/tools/biomuta/
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/
COSMIC	http://cancer.sanger.ac.uk/cosmic/
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
GWAS Catalog	https://www.ebi.ac.uk/gwas/
ICGC	https://dcc.icgc.org/
SwissVar	http://www.expasy.org/swissvar
humsva	http://www.uniprot.org/docs/humsva
TCGA	http://cancergenome.nih.gov/
UniProtKB	http://www.uniprot.org/

PTM Databases	Database URLs
dbPTM	http://dbPTM.mbc.nctu.edu.tw/
iPTMnet	http://proteininformationresource.org/iPTMnet/
O-GLYBASE	http://www.cbs.dtu.dk/databases/OGLYCBASE/
PHOSIDA	http://www.phosida.com/
Phospho.ELM	http://phospho.elm.eu.org/
PhosphoSitePlus	http://www.phosphosite.org/homeAction.action/
UniCarbKB	http://www.unicarbkb.org/
UniProtKB	http://www.uniprot.org/

2.1.1 GWAS Catalog

The NHGRI-EBI GWAS Catalog [86] is a manually curated literature-derived resource related to genome-wide association studies. It is a collaborative project between National Human Genome Research Institute (NHGRI), National Center for Biotechnology Information (NCBI), and the European Bioinformatics Institute (EBI). As of May 2016, the catalog contains over 100,000 SNPs with more than 20,000 SNP-trait associations, along with other GWAS-related statistical information, such as p-value, odds ratio, and confidence intervals.

2.1.2 ClinVar

ClinVar [87] offers freely available data about medically important variants and phenotypes. Each ClinVar submission also provides information on the relationship of the specific variation to human health, along with any evidence supporting that variation.

This database is reciprocally cross-referenced to dbSNP and dbVar, which maintain non-redundant information on SNP and structural variation, respectively. ClinVar also includes the phenotypic descriptions maintained in MedGen (<http://www.ncbi.nlm.nih.gov/medgen/>). The content in ClinVar can be divided into five major categories: submitter, variation, phenotype, interpretation, and evidence. The combinations of submitter, variation, and phenotype are given a unique accession number in the SCV000000000.0 (SCV) format. The data in ClinVar flow from dbSNP with some annotations from OMIM (<http://omim.org/>) and GeneReviews [88], and are maintained and released weekly. ClinVar accepts direct submissions following the process described at <http://www.ncbi.nlm.nih.gov/clinvar/docs/submit/> [87, 89].

2.1.3 UniProtKB/ SwissVar

SwissVar [90] is a portal to search genetic variants in UniProtKB/Swiss-Prot entries. Since UniProtKB/SwissProt contains the collection of human proteins along with detailed functional annotations, the SwissVar portal and the humsavar files (<http://www.uniprot.org/docs/humsavar>) provide access to this vast amount of information on protein variants and their relationship with disease [89, 90].

2.1.4 dbSNP

Since the inception of dbSNP [8] in 1998, it has been the central, public repository for genetic variations. The database, established by NCBI to serve as a general catalog of human SNPs and rare variations, provides information required for large-scale association studies, gene mapping, and evolutionary biology. The information in the dbSNP database is cross-referenced to relevant NCBI resources [8], such as Genbank, PubMed, LocusLink, and Human Genome Project data. All the information in dbSNP is freely available and can be downloaded in a variety of formats. The 1000 Genomes Project is a separate project aimed at building a platform wherein the association of human genetic variations to diseases is characterized with respect to their geographical and functional aspects. In this project, the genomes of 1092 individuals from 14 populations were constructed using whole-genome and whole-exome sequencing (WGS and WES, respectively). It provides a haplotype map of 38 million SNPs, 1.4 million short insertions and deletions, and more than 14,000 larger deletions, capturing over 95 % of the SNPs with greater than 1 % frequency in related populations. The 1000 Genomes Project includes ~50 %, 98 %, and 99.7 % of SNPs with ~0.1 %, 1.0 %, and 5.0 % frequencies, respectively, in ~2500 UK-sampled genomes [91]. Data from this project are available through the dbSNP portal.

2.1.5 The Cancer Genome Atlas (TCGA) Project

TCGA (<http://cancergenome.nih.gov/>) is a collaboration between the National Cancer Institute (NCI) and NHGRI being the most ambitious cancer sequencing project to date. A pilot project was initiated in 2006 with the aim of generating complete maps

of key genomic changes in major types and subtypes of cancer. Following the success of the pilot project, the NIH extended the research to include collection and characterization of additional tumor types. Samples from over 11,000 patients across 33 cancer types and subtypes, including 10 rare cancers, have now been characterized by TCGA consortium members. In addition to mutation calls, available data types include WGS, WES, RNA-Seq, and microarray data for mRNA and miRNA, methylation calls, copy number variations (CNVs), and clinical data. The project is officially coming to a close with the end of the 2016 fiscal year, but the resources generated and the impact on cancer research will continue to influence the field for years to come.

2.1.6 *The International Cancer Genome Consortium (ICGC)*

The ICGC [92] is an ongoing project aimed at the generation of extensive cataloging of genomic variations. Data collected include somatic and epigenetic modifications across 50 different cancer types and subtypes of global clinical and societal significance [92].

2.1.7 *Catalog of Somatic Mutations in Cancer (COSMIC)*

COSMIC [93] contains curated data from scientific papers about somatic mutations of human cancer and experimental data from the Cancer Genome Project at the Sanger Institute. COSMIC was released in 2014 with mutation data on four genes: BRAF, HRAS, KRAS2, and NRAS [94]. COSMIC is regularly updated, with the current build, V76, including 3,942,175 coding mutations, referencing 22,844 papers.

2.1.8 *BioMuta*

BioMuta is an integrated sequence feature database that includes curated cancer variation and disease association data wherein variations are mapped to genome, protein, and gene level annotations. BioMuta integrates somatic mutation and cancer data from various data sources including TCGA, ICGC, COSMIC, ClinVar, IntOGen [95], CSR (<https://hive.biochemistry.gwu.edu/dna.cgi?cmd=csr>) and UniProt, and augments the resulting catalog with manual curation of cancer-associated variations from publications [85]. All cancer types are re-assigned a disease ontology (DO) term to facilitate seamless integration of data from different sources and a harmonized approach to resulting pan-cancer analysis [96]. A total of 1,852,570 somatic SNVs were found to be associated with at least one cancer type in BioMuta v3.0 [85].

2.2 *Sequence-Specific PTM Annotations*

In addition to NGS approaches, high-throughput mass spectrometry studies have generated huge volumes of data on PTMs. Some databases which document the findings for a specific PTM include: Phospho.ELM [97], PhosphoSitePlus [98], PHOSIDA [99], and PhosPhAT [100] for phosphorylation sites; and O-GLYBASE [101] for glycosylation information. It is important to note that some PTM databases may contain information about all possible sites, exclusively predicted sites based on some underlying

molecular feature, or exclusively experimentally verified sites. Some of the most popular databases that maintain information across several types of PTMs are as follows:

2.2.1 *dbPTM*

The dbPTM database [102] compiles the information from publicly available databases like UniProtKB/SwissProt, Phospho.ELM, PhosphoSitePlus, O-GLYCBASE, dbSNO for S-nitrosylation [103], SysPTM [104], and the Human Protein Reference Database (HPRD) [105]. dbPTM then generates a dataset which contains only experimentally verified data, including manually curated data from research articles [106].

2.2.2 *UniProtKB*

As mentioned above, this is a publicly available resource with a wealth of diverse information, including information about PTMs and their corresponding functional and structural annotations [107, 108].

2.2.3 *UniCarbKB*

UniCarbKB [109, 110] aims at providing an extensive repository of curated data on glycan structures of glycoproteins, pathways and networks involved in glycosylation, and glycol-mediated processes. In the first release, UniCarbKB included 598 protein glycosylation sites including 35 glycoproteins, 502 structures, and 60 publications [109].

2.2.4 *iPTMnet*

iPTMnet is a bioinformatics database with information about several PTMs including phosphorylation, glycosylation, and acetylation, among others. It aims to bring together various bioinformatics and visualization tools, systems, and analyses used for text mining and data mining, databases, and ontologies into a single, unified platform to identify current gaps in knowledge and better explore PTM networks.

3 Methods

While the generation of data evidencing the existence of nsSNVs and PTMs is not a trivial task, there is a wealth of literature reviewing the various array-based [111] and NGS approaches [112] that have contributed to the growing wealth of data housed in the sources listed above. The method described herein encompasses a bioinformatics approach to analyze a collection of such data once it already exists. Taking into consideration the caveats that accompany the large-scale interpretation of such data, the following are the major steps for determining the impact of nsSNV on PTM sites:

3.1 *Pool Annotations Across Available Sources for Both nsSNVs and PTMs*

Retrieve relevant annotations from as many databases as fit the scope of your study to enable the most comprehensive study possible. For example purposes, consider a study aiming to characterize the impact of mutation on phosphorylation sites. With respect to nsSNV, you may consider retrieving data from any or all of the

following: dbSNP, COSMIC, UniProtKB, ICGC, IntOGen, and TCGA. For phosphorylation site data, you may want to consider all sources, or you may want to limit your retrieval of data to PHOSIDA due to their exclusion of data that has no experimental verification (predicted sites) or select data from UniProtKB based on evidence annotations. While the scope of data can always be winnowed down at a later point, it is generally beneficial to begin with a well-defined study design to avoid problems with data quality and incongruity.

3.2 Filter Each Dataset for Uniqueness and Quality Criteria

A major challenge of data integration is the rampant heterogeneity, both with respect to individual data points and to the representation of the same data, between different sources: if your dataset for a given feature is derived from more than one source, it is likely that there will be duplicate entries that may or may not be represented identically (*see Note 7*). Setting rational criteria applicable to all data sources is important for ensuring harmonization of the data. Distinct data points should be defined such that the resulting dataset is a union of all data unique to a specific source plus a collection of single data points representative of all data in the intersection of all sources. Because the presence or absence of a record in a given data source may be in itself an observation of interest, it may be advantageous to maintain the source data in a column even when a single record has more than one contributing source. If there is any specific quality criteria to impose upon the data, this is the recommended stage for applying such filters.

3.3 Map Relevant Features Between the Two Datasets

Depending on the specific study, there may be a number of features for which a correspondence map needs to be derived. In the most basic case, the minimum requirement will involve mapping the annotations from all sources from both features to a single positional reference frame. It is suggested that features be mapped to the position of the reference protein belonging to the complete human proteome in UniProtKB/SwissProt. This serves a dual purpose by limiting your result space to a set of confidently annotated proteins and facilitating simple downstream retrieval of Gene Ontology (GO) terms for functional enrichment analysis. In more complex cases, correspondence mapping could be used to prioritize features based on their functionality: for example, if we map SNVs to functional site annotations, we may choose to focus on SNVs that occur at an annotated functional site due to the presumed importance of that site to normal biological function.

3.4 Generate a New Dataset from the Overlapping Annotations

The new working dataset should be a table where rows are unique combinations of columns including position of overlap of the nsSNV and PTM sites, reference nucleotide, variant nucleotide, reference amino acid, variant amino acid, and evidence supporting the existence of the PTM or nsSNV. Additional columns can be maintained as relevant to the specific study design.

3.5 Count and Characterize the Data

Overlapping nsSNV and PTM sites can now be organized by many schemas, including but not limited to specific amino acid substitution, class of amino acid substitution, or expected gain or loss of PTMs. Grouping the data in such a way allows for discovery of any feature which may be disproportionately abundant or absent in the dataset. Enrichment analysis is conducted based on the abundance of a given feature with respect to the dataset and compared to the corresponding representation of that same feature in the entire human genome or proteome (depending on the specific case). This allows comparison of the expected ratio as obtained by the frequency of the feature in the genome/proteome to the observed ratio in the dataset, with statistical significance reported as a p-value ascertained by a cumulative binomial test [113]. For example, we can look at the enrichment of cancer-associated genes with mutation in active sites by calculating an expectation from the proportion of cancer-associated genes in the entire genome and comparing this to the observed frequency among the set of all genes with mutation in active sites.

3.6 Perform Functional Analysis

Functional analysis can be achieved computationally by a methodologically similar enrichment analysis of functional GO terms, or through the application of specific software like Ingenuity Pathway Analysis (IPA, www.qiagen.com/ingenuity) and PANTHER gene list analysis [114]. If desired, functional analysis and validation can also be performed using traditional wet-lab approaches like insertional mutagenesis.

3.7 Cross-Reference Existing Literature

Literature review is the backbone of any substantial research effort. Literature should be consulted to determine how much known PTM variability can be explained by nsSNV, and which, if any, of the previously undescribed PTM sites could be implicated in disease if the nsSNV were to alter the PTM status at the site of interest. Text mining approaches such as DiMeX can be used to extract such mutation–disease associations [115].

Although this approach has been described here specifically with application to assessing the impact of nsSNV on PTM in human disease, the same basic approach can be modified for assessing the impact of any feature of interest on another feature, as long as the features can be mapped by some underlying characteristic. This method has successfully been applied to a comparative analysis of N-linked glycosylation sites in eukaryotes [22], to a look at the impact of nsSNV on annotated human enzyme active sites [117], and to an integrated consideration of mutations across multiple features to determine functional profiles of distinct cancers [106]. Furthermore, while the mutation landscape with respect to disease associations has been well-characterized, we can substitute the nsSNV data for expression data to move toward a more comprehensive understanding of how aberrant regulation of a gene

may contribute to disease phenotypes. Some popular resources for expression data include BGee [118], BioXpress [119], and the Genotype-Tissue Expression Project (GTEx) [120].

4 Notes

An extensive literature review integrating the results of many analyses following the aforementioned protocol, among others, has been conducted to ascertain the impact of nsSNVs on different types of PTMs in the human proteome. It is estimated that the human genome has 25,000 genes [121] and over 1 million proteins [122] when accounting for various isoforms. Besides alternative splicing, PTMs are the major source of protein diversity. dbPTM stores the non-redundant data for 200 PTMs, and more than 66 % of PTM sites are present in the functional domain of proteins, suggesting that PTMs play a central role in regulation of protein interactions and functions [123]. Since protein function largely depends on a small number of critical sites [106], any dysfunction due to germline or somatic variations at these key PTM sites can potentially lead to a diseased condition.

In this chapter, we have summarized the impact of germline and somatic nsSNVs on different types of PTMs:

1. There are 77,734 experimentally verified phosphorylation sites reported in the human proteome, out of which 9383 sites were found to be replaced by amino acids that cannot be phosphorylated. 5466 of the sites are nsSNVs and not documented in dbSNP [106]. 6063, 2033, and 1283 variants resulted in a loss of phosphorylation at Ser, Thr, and Tyr residues, respectively [106]. Several important pathways like transcription, translation, cell cycle, and signal transduction are regulated by phosphorylation such that dysfunction caused by somatic or germline variations at the phosphorylation sites have been implicated in many diseases. Deregulation of kinase and phosphatase functions have been involved in several classes of human cancers [124], making kinase inhibitors important drug targets [125].

Different classes of phosphorylation-related enzymes seem to have related but different functional outcomes. Serine/threonine kinases constitute a large family of enzymes that are generally involved in cell signaling and have been implicated in several human cancers. Tyrosine kinases and their substrates are also important molecular cancer markers and attractive anticancer drug targets. About 518 protein kinase genes have been identified in the human genome, almost 100 of which are tyrosine kinases: more than 50 % of tyrosine kinases are linked to human cancers. Phosphorylation of tyrosine residues is a key

regulatory step in a series of cellular events including growth and proliferation. A large family of cell surface receptors with intrinsic tyrosine kinase activity is stimulated by growth factors like epidermal growth factor (EGF), platelet-derived growth factor (PDGF), vascular endothelial growth factor (VEGF), and fibroblast growth factor (FGF), all of which have been found involved in human cancers. For example, BCR-ABL tyrosine kinase is linked to chronic myelogenous leukemia (CML), ErbB2 receptor tyrosine kinase to breast cancer and, platelet-derived growth factor receptor tyrosine kinase to gliomas [125].

As described in the methods section above, pathway analysis and GO enrichment analysis can be performed to gain a deeper insight into the impact of nsSNVs that resulted in the loss of PTM sites. With respect to phosphorylation, previous analysis of this type reported angiogenesis and VEGF signaling to be significantly over-represented as compared to 21 prominent cancer-related pathways [106]. GO enrichment analysis for the same set of genes produced 32 biological processes, 23 molecular function terms, and 9 cellular component terms with over- or under-representation with respect to frequency of nsSNV. Biological processes represented include metabolic, cellular processes, cell cycle, and nucleobase-containing metabolic processes, and molecular function terms included kinase activity and nuclease binding.

2. Among the 16,170 experimentally verified N-linked glycosylation sites in the human proteome are 4372 nsSNVs affecting NXS/T motifs [52]. Of the affected N-linked glycosylation sites, 2375 somatic nsSNVs were not found in dbSNP at the time of analysis [106]. Interestingly, Ser/Thr motifs are disrupted by 2387 variations, direct changes in asparagine were caused by 1938 nsSNVs, and a change in proline caused the loss of 70 N-linked glycosylation sites [106]. The impact of nsSNVs was observed to be greater for N-linked glycosylation than for phosphorylation [124]. N-linked glycosylation has been implicated in several cancers including hepatocellular carcinoma (HCC). HCC is among the top ten most prevalent cancers in the world, with more than 80 % prevalence in developing countries, and is in the top five for cancer-related deaths worldwide [126]. One isoform of alpha-fetoprotein, AFP-L3, has been identified as one important marker for HCC [119, 127, 128]. A detailed analysis of the underlying molecular mechanism implicated the fucosylation of N-glycans as a possible cause for the secretion of hepatocyte-derived proteins (like AFP) into bile ducts instead of circulation in HCC patients [128]. In addition to AFP, elevated levels of biomarker GP73, a known target of N-glycosylation [127], in the sera of HCC

patients is also linked to defects in the secretory system of hepatocyte cells [129]. Thus, N-glycosylation can be used to identify potential biomarkers for hepatocellular carcinoma.

Pathway analysis emphasized two pathways which are most likely to be affected when variations result in loss of N-linked glycosylation sites: cadherin signaling pathway and Wnt signaling pathway. GO enrichment analysis of these pathways found 46 terms matched with different biological processes such as cell adhesion, nervous system development, and ectoderm development, meaning these broader processes are likely to be impacted by variations in N-glycosylation sites in these pathways [106].

3. O-linked glycosylation is generally linked with protein processing, mucin biosynthesis, proteoglycan core formation, and blood group proteins [130, 131]. In the human proteome, there are 2549 O-linked glycosylation sites out of which 205 are impacted by nsSNV. Among these disrupted O-glycosylation sites, 97 nsSNVs are not reported by dbSNP. 2150 of the 2549 sites correspond to the most studied *O*-*N*-acetylglucosamine (O-GalNAc) linkage, 138 of which are affected by nsSNVs. Out of 335 *O*-*N*-acetylglucosamine (O-GlcNAc) sites, 51 are altered by nsSNVs. It has been observed that O-linked glycosylation, as compared to N-linked glycosylation, has been inefficiently represented in terms of the number of variants at the site of O-glycosylation and the number of proteins actually impacted [106]. Like phosphorylation, O-GlcNAc cycles rapidly on protein and sometimes competes with phosphorylation for the same protein site [49, 132]. Glycomic studies, which attempt to define the total glycan content of a proteome, have revealed that O-GlcNAcylation shows significant cross talk with phosphorylation, resulting in altered regulation of transcription and signaling, as well as cytoskeleton function. O-GlcNAcylation also regulates a number of oncogenic proteins and tumor suppressor proteins. O-GlcNAcylation is most apparent in the brain and in the pancreas, thus explaining the link between deregulated O-GlcNAcylation and insulin resistance and glucose toxicity in diabetes and neurodegenerative disorders like Alzheimer's disease (AD) and Parkinson's disease [133].
4. Of 22,549 experimentally verified ubiquitination sites in the human proteome, 2055 of these sites were impacted by nsSNVs, 1214 of which were not found in dbSNP [106]. Ubiquitination is generally associated with degradation of proteins, but also plays key roles in many regulatory processes in a proteasome-independent fashion, including regulation of transcription, DNA repair, subnuclear trafficking, and endocytosis [134]. Protein stability regulation and the ubiquitin-protea-

some pathway is critical for understanding the basis of carcinogenesis. It has been observed that many proteins that are studied in breast and ovarian cancers are involved in ubiquitin pathways, including cyclins, CDK inhibitors, and SCF in cell cycle control, the breast and ovarian cancer suppressor BRCA1-BARD1, ErbB2/HER2/Neu and its ubiquitin ligase c-Cbl or CHIP, and estrogen receptors and downstream target Efp [134].

5. The human proteome contains 8253 acetylation sites, 863 of which have been estimated to lose their acetylation sites due to nsSNV. 512 of these potential nsSNVs have not been documented by dbSNP. Nt-acetylation plays a regulatory role in biological processes like protein degradation [67] and translocation of proteins [69]. Lysine acetylation plays a similar role in signaling, transcription, and stability of proteins [71]. Of 2194 N-acetylation sites, 201 are affected by nsSNVs, but it has been estimated that N6-acetyllysine has 6065 sites out of which 673 are altered by nsSNVs. One study reported that the loss of acetylation sites due to both germline and somatic nsSNVs is potentially less abundant than expected [135]. Additional research was done to compare the impact of acetylation on nsSNVs that are documented in dbSNP and those not present in dbSNP. Interestingly, these two subsets of nsSNVs revealed that the impact of nsSNVs in dbSNP on N6-acetyllysine sites is far less than N-acetylation sites in comparison with the somatic variations.
6. Of 680 experimentally verified methylation sites in the human proteome, 224 sites are affected by nsSNVs, 163 of which were not found in dbSNP. In the same study, 388 sites of methylarginine and 287 sites of methyllysine were observed: 172 arginine sites and 52 lysine sites were mutated by nsSNVs [106]. Research on methylation sites has shown that loss of methylation has significant correlation with cancer. For example, diffuse intrinsic pontine gliomas (DIPGs) are linked to the missense mutations of Lys27Met (K27 M) and Gly34Arg/Val (G34R/V) in the genes encoding histones H3.3 (H3F3A) and H3.1 (HIST3H1B) [83]. Mutation of K27 M lowers the overall methylation of H3K27me3 which inhibits the enzymatic activity of polycomb repressive complex 2 (PRC2), a pivotal player in the developmental regulation of gene expression. Methylation plays an important role in gene silencing in noncoding regions of the genome such as heterochromatin, which is transcriptionally inactive and extensively methylated as well. These highly methylated regions help to protect the genome from viral sequences and prevent them from integrating into the host genome. In contrast, the promoter regions in euchromatin are unmethylated except in mammalian cells where methylation is primarily found in cytosine-guanosine

(CpG) dinucleotides islands. These CpG islands are found near promoter regions and thus help in the inactivation of transcription [136–138]. Hypermethylation in tumor suppressor genes has been linked to many cancers, including hemopoietic malignancies. For example, the *MLH1* mismatch-repair gene found in colorectal cancer, *VHL* (von Hippel-Lindau) gene in renal cancer, and *BRCA1* in early breast cancer [137, 139] have all been reported to be hypermethylated in disease states. In some tumors, genes may remain silenced due to hypermethylation, such as the gene for O⁶-methylguanine-DNA methyltransferase (*MGMT*) involved in DNA repair [140] and cyclin-dependent kinase inhibitor 2B (*CDKN2B*) that encodes for cell-cycle regulator p15.

5 Conclusions

nsSNVs are changes in the nucleotide sequence that result in the translation of an amino acid sequence different than that of the original non-variant protein product. When these changes occur at positions normally subject to PTMs, downstream effects can range from silent to drastic, benevolent to deleterious. Specifically, nsSNV can result in the gain or loss of a PTM by modifying the codon for the attachment site of a given PTM or by altering sequons that are important for enzyme recognition. Although many examples in this chapter and throughout literature focus on the loss of PTM in disease, it is important to note that gain of PTM can have both beneficial and harmful consequences, just like loss of PTM. Change in PTMs have been associated with a variety of diseases including diabetes, neurodegenerative disorders like Alzheimer's disease and Parkinson's disease, and several classes of cancer. Customization of the general protocol provided herein can promote further discovery of variation-impacted PTM and any resulting disease associations. Furthermore, this method can be adapted to an expression-based survey of PTM-affected proteins with a few minor modifications.

Despite the availability of data for both nsSNVs and PTMs, there are many challenges to and important considerations regarding the proposed analysis. First and foremost, increases in the volume of sequence data generated by NGS methods, enhanced by a layer of manual curation, has increased the complexity of data handling. The vocabulary used among different databases is also fragmented, which presents challenges to cross-referencing and harmonization of data. The field is aware of these discrepancies and several ongoing efforts intend to bridge the knowledge and interoperability gaps across and between databases, large-scale collaboration efforts, the biocuration community, policy enforcers, and the technology providers. Additional efforts are being made to address the emerging need of standardization of all genomic data and metadata.

One such initiative, GlycoRDF, is a global effort among bioinformatics scientists and glycobiologists to develop a Resource Description Framework (RDF) for glycomics data with an emphasis on sequence, biological source, publication, and experimental data [141]. (For more information on GlycoRDF, please visit <http://www.glycoinfo.org/GlycoRDF/>.) Another interesting effort is that of the Protein Standards Initiative on Protein Modification (PSI-MOD), which includes members from proteomics, mass spectrometry, and bioinformatics communities and aims to define proteomics standards to facilitate confident data comparison, annotation, and validation of entries across the different databases [142]. A disease-centric initiative, the ICGC-TCGA DREAM Somatic Mutation Calling –Tumor Heterogeneity Challenge (SMC-Het), comes in the form of a challenge posed to the broader scientific community with the aim of ranking and improving HTS analysis algorithms to enable the reliable identification of somatic mutations. (Additional information can be accessed at <https://www.synapse.org/#!/Synapse:syn2813581/wiki/303137>). Thus, while the relatively rapid propagation of variation and PTM data in the early 2000s resulted in a number of disjointed and highly heterogeneous data repositories, there is a general consensus among diverse stakeholders, from government agencies to clinicians to the research community at large, that development of data standardization, harmonization, and validation practices will greatly enhance the impact of this type of information on public health.

References

1. Crick F (1970) Central dogma of molecular biology. *Nature* 227(5258):561–563. doi:10.1038/227561a0
2. Koonin EV (2012) Does the central dogma still stand? *Biol Direct* 7:27. doi:10.1186/1745-6150-7-27
3. Chesebro B (2003) Introduction to the transmissible spongiform encephalopathies or prion diseases. *Br Med Bull* 66:1–20
4. Chien P, Weissman JS, DePace AH (2004) Emerging principles of conformation-based prion inheritance. *Annu Rev Biochem* 73:617–656. doi:10.1146/annurev.biochem.72.121801.161837
5. Munch C, Bertolotti A (2012) Propagation of the prion phenomenon: beyond the seeding principle. *J Mol Biol* 421(4–5):491–498. doi:10.1016/j.jmb.2011.12.061
6. Sachidanandam R, Weissman D, SC S, JM K, LD S, Marth G, Sherry S, JC M, BJ M, DL W, SE H, CG C, PC C, CM R, Ning Z, Rogers J, DR B, PY K, ER M, RT Y, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, RH W, JD MP, Gilman B, Schaffner S, WJ VE, Reich D, Higgins J, MJ D, Blumenstiel B, Baldwin J, Stange-Thomann N, MC Z, Linton L, ES L, Altshuler D, International SNPMPWG (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409(6822):928–933. doi:10.1038/35057149
7. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* 33(4):518–521. doi:10.1038/ng1128
8. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29(1):308–311
9. Genomes Project C, GR A, Altshuler D, Auton A, LD B, RM D, RA G, ME H, GA

- MV (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073. doi:[10.1038/nature09534](https://doi.org/10.1038/nature09534)
10. Lehne B, Lewis CM, Schlitt T (2011) From SNPs to genes: disease association at the gene level. *PLoS One* 6(6):e20133. doi:[10.1371/journal.pone.0020133](https://doi.org/10.1371/journal.pone.0020133)
 11. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans DM, Leung HT, Marchini JL, Morris AP, Spencer CC, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop TD, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Matthew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop MG, Connell J, Dominiczak A, Marcano CA, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue M, Caulfield M, Farrall M, Barton A, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hilder SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DP, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JR, Shields BM, Weedon MN, Hattersley AT, Hitman GA, Walker M, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lyons E, Vannberg F, Hill AV, Bradbury LA, Farrar C, Pointon JJ, Wordsworth P, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Gough SC, Seal S, Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Rockett KA, Bumpstead SJ, Chaney A, Downes K, Ghorji MJ, Gwilliam R, Hunt SE, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Widdon C, Withers D, Cardin NJ, Ferreira T, Pereira-Gale J, Hallgrimsdottir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Mitchell SL, Newby PR, Brand OJ, Carr-Smith J, Pearce SH, Reveille JD, Zhou X, Sims AM, Dowling A, Taylor J, Doan T, Davis JC, Savage L, Ward MM, Learch TL, Weisman MH, Brown M (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* 39(11):1329–1337. doi:[10.1038/ng.2007.17](https://doi.org/10.1038/ng.2007.17)
 12. Levy A, Hall L, Yeudall WA, Lightman SL (1994) p53 gene mutations in pituitary adenomas: rare events. *Clin Endocrinol (Oxf)* 41(6):809–814
 13. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurler M, Korn JM, Kristiansson K, Lee C, McCarroll SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghorji MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52–58. doi:[10.1038/nature09298](https://doi.org/10.1038/nature09298)
 14. Eilbeck K, Lewis SE (2004) Sequence ontology annotation guide. *Comp Funct Genomics* 5(8):642–647. doi:[10.1002/cfg.446](https://doi.org/10.1002/cfg.446)
 15. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y (2012) Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* 13(Suppl 8):S8. doi:[10.1186/1471-2164-13-S8-S8](https://doi.org/10.1186/1471-2164-13-S8-S8)
 16. Varani G, McClain WH (2000) The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep* 1(1):18–23. doi:[10.1093/embo-reports/kvd001](https://doi.org/10.1093/embo-reports/kvd001)
 17. Weatherall DJ (2004) Thalassaemia: the long road from bedside to genome. *Nat Rev Genet* 5(8):625–631. doi:[10.1038/nrg1406](https://doi.org/10.1038/nrg1406)

18. Griffiths A (2000) An introduction to genetic analysis, 7th edn. W.H. Freeman, New York
19. Grotenbreg G, Ploegh H (2007) Chemical biology: dressed-up proteins. *Nature* 446(7139):993–995. doi:[10.1038/446993a](https://doi.org/10.1038/446993a)
20. Jenuwein T, Allis CD (2001) Translating the histone code. *Science* 293(5532):1074–1080. doi:[10.1126/science.1063127](https://doi.org/10.1126/science.1063127)
21. Bode AM, Dong Z (2004) Post-translational modification of p53 in tumorigenesis. *Nat Rev Cancer* 4(10):793–805. doi:[10.1038/nrcl455](https://doi.org/10.1038/nrcl455)
22. Lam PV, Goldman R, Karagiannis K, Narsule T, Simonyan V, Soika V, Mazumder R (2013) Structure-based comparative analysis and prediction of N-linked glycosylation sites in evolutionarily distant eukaryotes. *Genomics Proteomics Bioinformatics* 11(2):96–104. doi:[10.1016/j.gpb.2012.11.003](https://doi.org/10.1016/j.gpb.2012.11.003)
23. Khoury GA, Baliban RC, Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep* 1. doi:[10.1038/srep00090](https://doi.org/10.1038/srep00090)
24. Burnett G, Kennedy EP (1954) The enzymatic phosphorylation of proteins. *J Biol Chem* 211(2):969–980
25. Ubersax JA, Ferrell JE Jr (2007) Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol* 8(7):530–541. doi:[10.1038/nrm2203](https://doi.org/10.1038/nrm2203)
26. Cohen P (2000) The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem Sci* 25(12):596–601
27. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298(5600):1912–1934. doi:[10.1126/science.1075762](https://doi.org/10.1126/science.1075762)
28. Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* 27(10):514–520
29. Ciesla J, Fraczyk T, Rode W (2011) Phosphorylation of basic amino acid residues in proteins: important but easily missed. *Acta Biochim Pol* 58(2):137–148
30. Zisch AH, D'Alessandri L, Amrein K, Ranscht B, Winterhalter KH, Vaughan L (1995) The glypiated neuronal cell adhesion molecule contactin/F11 complexes with src-family protein tyrosine kinase Fyn. *Mol Cell Neurosci* 6(3):263–279. doi:[10.1006/mcne.1995.1021](https://doi.org/10.1006/mcne.1995.1021)
31. Wolfe BL, Trejo J (2007) Clathrin-dependent mechanisms of G protein-coupled receptor endocytosis. *Traffic* 8(5):462–470. doi:[10.1111/j.1600-0854.2007.00551.x](https://doi.org/10.1111/j.1600-0854.2007.00551.x)
32. Qureshi S, Galiveeti S, Bichet DG, Roth J (2014) Diabetes insipidus: celebrating a century of vasopressin therapy. *Endocrinology* 155(12):4605–4621. doi:[10.1210/en.2014-1385](https://doi.org/10.1210/en.2014-1385)
33. Rene P, Le Gouill C, Pogozheva ID, Lee G, Mosberg HI, Farooqi IS, Valenzano KJ, Bouvier M (2010) Pharmacological chaperones restore function to MC4R mutants responsible for severe early-onset obesity. *J Pharmacol Exp Ther* 335(3):520–532. doi:[10.1124/jpet.110.172098](https://doi.org/10.1124/jpet.110.172098)
34. Tzekov R, Stein L, Kaushal S (2011) Protein misfolding and retinal degeneration. *Cold Spring Harb Perspect Biol* 3(11):a007492. doi:[10.1101/cshperspect.a007492](https://doi.org/10.1101/cshperspect.a007492)
35. Butcher AJ, Prihandoko R, Kong KC, McWilliams P, Edwards JM, Bottrill A, Mistry S, Tobin AB (2011) Differential G-protein-coupled receptor phosphorylation provides evidence for a signaling bar code. *J Biol Chem* 286(13):11506–11518. doi:[10.1074/jbc.M110.154526](https://doi.org/10.1074/jbc.M110.154526)
36. Lappano R, Maggiolini M (2011) G protein-coupled receptors: novel targets for drug discovery in cancer. *Nat Rev Drug Discov* 10(1):47–60. doi:[10.1038/nrd3320](https://doi.org/10.1038/nrd3320)
37. Marth JD, Grewal PK (2008) Mammalian glycosylation in immunity. *Nat Rev Immunol* 8(11):874–887. doi:[10.1038/nri2417](https://doi.org/10.1038/nri2417)
38. Rudd PM, Elliott T, Cresswell P, Wilson IA, Dwek RA (2001) Glycosylation and the immune system. *Science* 291(5512):2370–2376
39. Marino K, Bones J, Kattla JJ, Rudd PM (2010) A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol* 6(10):713–723. doi:[10.1038/nchembio.437](https://doi.org/10.1038/nchembio.437)
40. Dwek RA (1996) Glycobiology: toward understanding the function of sugars. *Chem Rev* 96(2):683–720
41. Spiro RG (2002) Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* 12(4):43R–56R
42. Moremen KW, Tiemeyer M, Nairn AV (2012) Vertebrate protein glycosylation: diversity, synthesis and function. *Nat Rev Mol Cell Biol* 13(7):448–462. doi:[10.1038/nrm3383](https://doi.org/10.1038/nrm3383)
43. Ohtsubo K, Marth JD (2006) Glycosylation in cellular mechanisms of health and disease. *Cell* 126(5):855–867. doi:[10.1016/j.cell.2006.08.019](https://doi.org/10.1016/j.cell.2006.08.019)
44. Gagneux P, Varki A (1999) Evolutionary considerations in relating oligosaccharide

- diversity to biological function. *Glycobiology* 9(8):747–755
45. Pompach P, Brnakova Z, Sanda M, Wu J, Edwards N, Goldman R (2013) Site-specific glycoforms of haptoglobin in liver cirrhosis and hepatocellular carcinoma. *Mol Cell Proteomics* 12(5):1281–1293. doi:[10.1074/mcp.M112.023259](https://doi.org/10.1074/mcp.M112.023259)
 46. Landsteiner K (1931) Individual differences in human blood. *Science* 73(1894):403–409. doi:[10.1126/science.73.1894.403](https://doi.org/10.1126/science.73.1894.403)
 47. Shriver Z, Raguram S, Sasisekharan R (2004) Glycomics: a pathway to a class of new and improved therapeutics. *Nat Rev Drug Discov* 3(10):863–873. doi:[10.1038/nrd1521](https://doi.org/10.1038/nrd1521)
 48. Miura Y, Endo T (2016) Glycomics and glycoproteomics focused on aging and age-related diseases—Glycans as a potential biomarker for physiological alterations. *Biochim Biophys Acta*. doi:[10.1016/j.bbagen.2016.01.013](https://doi.org/10.1016/j.bbagen.2016.01.013)
 49. Varki A (2015) *Essentials of glycobiology*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY)
 50. Scott H, Panin VM (2014) The role of protein N-glycosylation in neural transmission. *Glycobiology* 24(5):407–417. doi:[10.1093/glycob/cwu015](https://doi.org/10.1093/glycob/cwu015)
 51. Aebi M, Bernasconi R, Clerc S, Molinari M (2010) N-glycan structures: recognition and processing in the ER. *Trends Biochem Sci* 35(2):74–82. doi:[10.1016/j.tibs.2009.10.001](https://doi.org/10.1016/j.tibs.2009.10.001)
 52. Mazumder R, Morampudi KS, Motwani M, Vasudevan S, Goldman R (2012) Proteome-wide analysis of single-nucleotide variations in the N-glycosylation sequon of human genes. *PLoS One* 7(5):e36212. doi:[10.1371/journal.pone.0036212](https://doi.org/10.1371/journal.pone.0036212)
 53. Trombetta ES (2003) The contribution of N-glycans and their processing in the endoplasmic reticulum to glycoprotein biosynthesis. *Glycobiology* 13(9):77R–91R. doi:[10.1093/glycob/cwg075](https://doi.org/10.1093/glycob/cwg075)
 54. Ramachandran R, Noorbakhsh F, Defea K, Hollenberg MD (2012) Targeting proteinase-activated receptors: therapeutic potential and challenges. *Nat Rev Drug Discov* 11(1):69–86. doi:[10.1038/nrd3615](https://doi.org/10.1038/nrd3615)
 55. Arora P, Ricks TK, Trejo J (2007) Protease-activated receptor signalling, endocytic sorting and dysregulation in cancer. *J Cell Sci* 120(Pt 6):921–928. doi:[10.1242/jcs.03409](https://doi.org/10.1242/jcs.03409)
 56. Zauner G, Kozak RP, Gardner RA, Fernandes DL, Deelder AM, Wuhrer M (2012) Protein O-glycosylation analysis. *Biol Chem* 393(8):687–708. doi:[10.1515/hsz-2012-0144](https://doi.org/10.1515/hsz-2012-0144)
 57. Wojcikiewicz RJ (2004) Regulated ubiquitination of proteins in GPCR-initiated signaling pathways. *Trends Pharmacol Sci* 25(1):35–41. doi:[10.1016/j.tips.2003.11.008](https://doi.org/10.1016/j.tips.2003.11.008)
 58. Alonso V, Friedman PA (2013) Minireview: ubiquitination-regulated G protein-coupled receptor signaling and trafficking. *Mol Endocrinol* 27(4):558–572. doi:[10.1210/me.2012-1404](https://doi.org/10.1210/me.2012-1404)
 59. Hammond-Martel I, Yu H, Affar el B (2012) Roles of ubiquitin signaling in transcription regulation. *Cell Signal* 24(2):410–421. doi:[10.1016/j.cellsig.2011.10.009](https://doi.org/10.1016/j.cellsig.2011.10.009)
 60. Chen ZJ, Sun LJ (2009) Nonproteolytic functions of ubiquitin in cell signaling. *Mol Cell* 33(3):275–286. doi:[10.1016/j.molcel.2009.01.014](https://doi.org/10.1016/j.molcel.2009.01.014)
 61. Norskov-Lauritsen L, Brauner-Osborne H (2015) Role of post-translational modifications on structure, function and pharmacology of class C G protein-coupled receptors. *Eur J Pharmacol* 763(Pt B):233–240. doi:[10.1016/j.ejphar.2015.05.015](https://doi.org/10.1016/j.ejphar.2015.05.015)
 62. Espinosa JM (2008) Histone H2B ubiquitination: the cancer connection. *Genes Dev* 22(20):2743–2749. doi:[10.1101/gad.1732108](https://doi.org/10.1101/gad.1732108)
 63. Van Damme P, Hole K, Pimenta-Marques A, Helsen K, Vandekerckhove J, Martinho RG, Gevaert K, Arnesen T (2011) NatF contributes to an evolutionary shift in protein N-terminal acetylation and is important for normal chromosome segregation. *PLoS Genet* 7(7):e1002169. doi:[10.1371/journal.pgen.1002169](https://doi.org/10.1371/journal.pgen.1002169)
 64. Starheim KK, Gevaert K, Arnesen T (2012) Protein N-terminal acetyltransferases: when the start matters. *Trends Biochem Sci* 37(4):152–161. doi:[10.1016/j.tibs.2012.02.003](https://doi.org/10.1016/j.tibs.2012.02.003)
 65. Yi CH, Pan H, Seebacher J, Jang IH, Hyberts SG, Heffron GJ, Vander Heiden MG, Yang R, Li F, Locasale JW, Sharfi H, Zhai B, Rodriguez-Mias R, Luithardt H, Cantley LC, Daley GQ, Asara JM, Gygi SP, Wagner G, Liu CF, Yuan J (2011) Metabolic regulation of protein N-alpha-acetylation by Bcl-xL promotes cell survival. *Cell* 146(4):607–620. doi:[10.1016/j.cell.2011.06.050](https://doi.org/10.1016/j.cell.2011.06.050)
 66. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM, Carey JC, Opitz JM, Stevens CA, Jiang T, Schank C, Fain HD, Robison R, Dalley B, Chin S, South ST, Pysker TJ, Jorde LB, Hakonarson H, Lillehaug JR, Biesecker LG, Yandell M, Arnesen T, Lyon GJ (2011) Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am*

- J Hum Genet 89(1):28–43. doi:[10.1016/j.ajhg.2011.05.017](https://doi.org/10.1016/j.ajhg.2011.05.017)
67. Hwang CS, Shemorry A, Varshavsky A (2010) N-terminal acetylation of cellular proteins creates specific degradation signals. *Science* 327(5968):973–977. doi:[10.1126/science.1183147](https://doi.org/10.1126/science.1183147)
 68. Forte GM, Pool MR, Stirling CJ (2011) N-terminal acetylation inhibits protein targeting to the endoplasmic reticulum. *PLoS Biol* 9(5):e1001073. doi:[10.1371/journal.pbio.1001073](https://doi.org/10.1371/journal.pbio.1001073)
 69. Scott DC, Monda JK, Bennett EJ, Harper JW, Schulman BA (2011) N-terminal acetylation acts as an avidity enhancer within an interconnected multiprotein complex. *Science* 334(6056):674–678. doi:[10.1126/science.1209307](https://doi.org/10.1126/science.1209307)
 70. Hofmann I, Munro S (2006) An N-terminally acetylated Arf-like GTPase is localised to lysosomes and affects their motility. *J Cell Sci* 119(Pt 8):1494–1503. doi:[10.1242/jcs.02958](https://doi.org/10.1242/jcs.02958)
 71. Sadoul K, Boyault C, Pabion M, Khochbin S (2008) Regulation of protein turnover by acetyltransferases and deacetylases. *Biochimie* 90(2):306–312. doi:[10.1016/j.biochi.2007.06.009](https://doi.org/10.1016/j.biochi.2007.06.009)
 72. Spange S, Wagner T, Heinzel T, Kramer OH (2009) Acetylation of non-histone proteins modulates cellular signalling at multiple levels. *Int J Biochem Cell Biol* 41(1):185–198. doi:[10.1016/j.biocel.2008.08.027](https://doi.org/10.1016/j.biocel.2008.08.027)
 73. Li QQ, Hao JJ, Zhang Z, Hsu I, Liu Y, Tao Z, Lewi K, Metwalli AR, Agarwal PK (2016) Histone deacetylase inhibitor-induced cell death in bladder cancer is associated with chromatin modification and modifying protein expression: A proteomic approach. *Int J Oncol*. doi:[10.3892/ijo.2016.3478](https://doi.org/10.3892/ijo.2016.3478)
 74. Hamey JJ, Winter DL, Yagoub D, Overall CM, Hart-Smith G, Wilkins MR (2016) Novel N-terminal and lysine methyltransferases that target translation elongation factor 1A in yeast and human. *Mol Cell Proteomics* 15(1):164–176. doi:[10.1074/mcp.M115.052449](https://doi.org/10.1074/mcp.M115.052449)
 75. Liu H, Galka M, Mori E, Liu X, Lin YF, Wei R, Pittock P, Voss C, Dhami G, Li X, Miyaji M, Lajoie G, Chen B, Li SS (2013) A method for systematic mapping of protein lysine methylation identifies functions for HP1beta in DNA damage response. *Mol Cell* 50(5):723–735. doi:[10.1016/j.molcel.2013.04.025](https://doi.org/10.1016/j.molcel.2013.04.025)
 76. Sayegh J, Webb K, Cheng D, Bedford MT, Clarke SG (2007) Regulation of protein arginine methyltransferase 8 (PRMT8) activity by its N-terminal domain. *J Biol Chem* 282(50):36444–36453. doi:[10.1074/jbc.M704650200](https://doi.org/10.1074/jbc.M704650200)
 77. Bedford MT, Richard S (2005) Arginine methylation an emerging regulator of protein function. *Mol Cell* 18(3):263–272. doi:[10.1016/j.molcel.2005.04.003](https://doi.org/10.1016/j.molcel.2005.04.003)
 78. Yang Y, Bedford MT (2013) Protein arginine methyltransferases and cancer. *Nat Rev Cancer* 13(1):37–50. doi:[10.1038/nrc3409](https://doi.org/10.1038/nrc3409)
 79. Byvoet P, Shepherd GR, Hardin JM, Noland BJ (1972) The distribution and turnover of labeled methyl groups in histone fractions of cultured mammalian cells. *Arch Biochem Biophys* 148(2):558–567
 80. Shi Y, Lan F, Matson C, Mulligan P, Whetstone JR, Cole PA, Casero RA, Shi Y (2004) Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* 119(7):941–953. doi:[10.1016/j.cell.2004.12.012](https://doi.org/10.1016/j.cell.2004.12.012)
 81. Greer EL, Shi Y (2012) Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet* 13(5):343–357. doi:[10.1038/nrg3173](https://doi.org/10.1038/nrg3173)
 82. Young NL, Dimaggio PA, Garcia BA (2010) The significance, development and progress of high-throughput combinatorial histone code analysis. *Cell Mol Life Sci* 67(23):3983–4000. doi:[10.1007/s00018-010-0475-7](https://doi.org/10.1007/s00018-010-0475-7)
 83. Lewis PW, Muller MM, Koletsky MS, Cordero F, Lin S, Banaszynski LA, Garcia BA, Muir TW, Becher OJ, Allis CD (2013) Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. *Science* 340(6134):857–861. doi:[10.1126/science.1232245](https://doi.org/10.1126/science.1232245)
 84. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 11(10):685–696. doi:[10.1038/nrg2841](https://doi.org/10.1038/nrg2841)
 85. Wu TJ, Shamsaddini A, Pan Y, Smith K, Crichton DJ, Simonyan V, Mazumder R (2014) A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). *Database (Oxford)* 2014:bau022. doi:[10.1093/database/bau022](https://doi.org/10.1093/database/bau022)
 86. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorf L, Parkinson H (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(Database issue):D1001–D1006. doi:[10.1093/nar/gkt1229](https://doi.org/10.1093/nar/gkt1229)
 87. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR

- (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42(Database issue):D980–D985. doi:[10.1093/nar/gkt1113](https://doi.org/10.1093/nar/gkt1113)
88. RA P, MP A, HH A et al (eds) (1993) *GeneReviews(R)*. University of Washington, Seattle (WA)
89. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B (2006) The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34(Database issue):D187–D191
90. Mottaz A, David FP, Veuthey AL, Yip YL (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* 26(6):851–852. doi:[10.1093/bioinformatics/btq028](https://doi.org/10.1093/bioinformatics/btq028)
91. Genomes Project C, GR A, Auton A, LD B, MA DP, RM D, RE H, HM K, GT M, GA MV (2012) An integrated map of genetic variation from 1092 human genomes. *Nature* 491(7422):56–65. doi:[10.1038/nature11632](https://doi.org/10.1038/nature11632)
92. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MM, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SO, Joly Y, Kato K, Kennedy KL, Nicolas P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clement B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayes M, Botwell DD, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, Lopez-Otin C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, Flicek P, Getz G, Guigo R, Guo G, Haussler D, Heath S, Hubbard TJ, Jiang T, Jones SM, Li Q, Lopez-Bigas N, Luo R, Muthuswamy L, Ouellette BF, Pearson JV, Puente XS, Quesada V, Raphael BJ, Sander C, Shibata T, Speed TP, Stein LD, Stuart JM, Teague JW, Totoki Y, Tsunoda T, Valencia A, Wheeler DA, Wu H, Zhao S, Zhou G, Stein LD, Guigo R, Hubbard TJ, Joly Y, Jones SM, Kasprzyk A, Lathrop M, Lopez-Bigas N, Ouellette BF, Spellman PT, Teague JW, Thomas G, Valencia A, Yoshida T, Kennedy KL, Axton M, Dyke SO, Futreal PA, Gerhard DS, Gunter C, Guyer M, Hudson TJ, McPherson JD, Miller LJ, Ozenberger B, Shaw KM, Kasprzyk A, Stein LD, Zhang J, Haider SA, Wang J, Yung CK, Cros A, Liang Y, Gnaneshan S, Guberman J, Hsu J, Bobrow M, Chalmers DR, Hasel KW, Joly Y, Kaan TS, Kennedy KL, Knoppers BM, Lowrance WW, Masui T, Nicolas P, Rial-Sebbag E, Rodriguez LL, Vergely C, Yoshida T, Grimmond SM, Biankin AV, Bowtell DD, Cloonan N, deFazio A, Eshleman JR, Etemadmoghadam D, Gardiner BB, Kench JG, Scarpa A, Sutherland RL, Tempero MA, Waddell NJ, Wilson PJ, McPherson JD, Gallinger S, Tsao MS, Shaw PA, Petersen GM, Mukhopadhyay D, Chin L, DePinho RA, Thayer S, Muthuswamy L, Shazand K, Beck T, Sam M, Timms L, Ballin V, Lu Y, Ji J, Zhang X, Chen F, Hu X, Zhou G, Yang Q, Tian G, Zhang L, Xing X, Li X, Zhu Z, Yu Y, Yu J, Yang H, Lathrop M, Tost J, Brennan P, Holcatova I, Zaridze D, Brazma A, Egevard L, Prokhorouchouk E, Banks RE, Uhlen M, Cambon-Thomsen A, Viksna J, Ponten F, Skryabin K, Stratton MR, Futreal PA, Birney E, Borg A, Borresen-Dale AL, Caldas C, Foekens JA, Martin S, Reis-Filho JS, Richardson AL, Sotiriou C, Stunnenberg HG, Thoms G, van de Vijver M, van't Veer L, Calvo F, Birnbaum D, Blanche H, Boucher P, Boyault S, Chabannon C, Gut I, Masson-Jacquemier JD, Lathrop M, Pauporte I, Pivot X, Vincent-Salomon A, Tabone E, Theillet C, Thomas G, Tost J, Treilleux I, Calvo F, Bioulac-Sage P, Clement B, Decaens T, Degos F, Franco D, Gut I, Gut M, Heath S, Lathrop M, Samuel D, Thomas G, Zucman-Rossi J, Lichter P, Eils R, Brors B, Korbel JO, Korshunov A, Landgraf P, Lehrach H, Pfister S, Radlwimmer B, Reifemberger G, Taylor MD, von Kalle C, Majumder PP, Sarin R, Rao TS, Bhan MK, Scarpa A, Pederzoli P, Lawlor RA, Delledonne M, Bardelli A, Biankin AV, Grimmond SM, Gress T, Klimstra D, Zamboni G, Shibata T, Nakamura Y, Nakagawa H, Kusada J, Tsunoda T, Miyano S, Aburatani H, Kato K, Fujimoto A, Yoshida T, Campo E, Lopez-Otin C, Estivill X, Guigo R, de Sanjose S, Piris MA, Montserrat E, Gonzalez-Diaz M, Puente XS, Jares P, Valencia A, Himmelbauer H, Quesada V, Bea S, Stratton MR, Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver

- M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Borresen-Dale AL, Caldas C, Foekens JA, Stunnenberg HG, van't Veer L, Easton DF, Spellman PT, Martin S, Barker AD, Chin L, Collins FS, Compton CC, Ferguson ML, Gerhard DS, Getz G, Gunter C, Guttmacher A, Guyer M, Hayes DN, Lander ES, Ozenberger B, Penny R, Peterson J, Sander C, Shaw KM, Speed TP, Spellman PT, Vockley JG, Wheeler DA, Wilson RK, Hudson TJ, Chin L, Knoppers BM, Lander ES, Lichter P, Stein LD, Stratton MR, Anderson W, Barker AD, Bell C, Bobrow M, Burke W, Collins FS, Compton CC, DePinho RA, Easton DF, Futreal PA, Gerhard DS, Green AR, Guyer M, Hamilton SR, Hubbard TJ, Kallioniemi OP, Kennedy KL, Ley TJ, Liu ET, Lu Y, Majumder P, Marra M, Ozenberger B, Peterson J, Schafer AJ, Spellman PT, Stunnenberg HG, Wainwright BJ, Wilson RK, Yang H (2010) International network of cancer genome projects. *Nature* 464(7291):993–998. doi:10.1038/nature08987
93. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 39(Database issue):D945–D950. doi:10.1093/nar/gkq929
94. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* 91(2):355–358. doi:10.1038/sj.bjc.6601894
95. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* 10(11):1081–1082. doi:10.1038/nmeth.2642
96. Wu TJ, Schriml LM, Chen QR, Colbert M, Crichton DJ, Finney R, Hu Y, Kibbe WA, Kincaid H, Meerzaman D, Mitraka E, Pan Y, Smith KM, Srivastava S, Ward S, Yan C, Mazumder R (2015) Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database (Oxford)* 2015:bav032. doi:10.1093/database/bav032
97. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res* 39(Database issue):D261–D267. doi:10.1093/nar/gkq1104
98. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40(Database issue):D261–D270. doi:10.1093/nar/gkr1122
99. Gnad F, Ren S, Cox J, Olsen JV, Macek B, Orosi M, Mann M (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8(11):R250. doi:10.1186/gb-2007-8-11-r250
100. Heazlewood JL, Durek P, Hummel J, Selbig J, Weckwerth W, Walther D, Schulze WX (2008) PhosphAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res* 36(Database issue):D1015–D1021. doi:10.1093/nar/gkm812
101. Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 27(1):370–372
102. Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res* 34(Database issue):D622–D627. doi:10.1093/nar/gkj083
103. Lee TY, Chen YJ, Lu CT, Ching WC, Teng YC, Huang HD, Chen YJ (2012) dbSNO: a database of cysteine S-nitrosylation. *Bioinformatics* 28(17):2293–2295. doi:10.1093/bioinformatics/bts436
104. Li J, Jia J, Li H, Yu J, Sun H, He Y, Lv D, Yang X, Glocker MO, Ma L, Yang J, Li L, Li W, Zhang G, Liu Q, Li Y, Xie L (2014) SysPTM 2.0: an updated systematic resource for post-translational modification. *Database (Oxford)* 2014:bau025. doi:10.1093/database/bau025
105. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human protein reference database--2009 update. *Nucleic Acids Res* 37(Database issue):D767–D772. doi:10.1093/nar/gkn892

106. Pan Y, Karagiannis K, Zhang H, Dingerdissen H, Shamsaddini A, Wan Q, Simonyan V, Mazumder R (2014) Human germline and pan-cancer variomes and their distinct functional profiles. *Nucleic Acids Res* 42(18):11570–11588. doi:[10.1093/nar/gku772](https://doi.org/10.1093/nar/gku772)
107. Farriol-Mathis N, Garavelli JS, Boeckmann B, Duvaud S, Gasteiger E, Gateau A, Veuthey AL, Bairoch A (2004) Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* 4(6):1537–1550. doi:[10.1002/pmic.200300764](https://doi.org/10.1002/pmic.200300764)
108. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 32(Database issue):D115–D119
109. Campbell MP, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita KF, Lisacek F, Packer NH (2014) UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res* 42(Database issue):D215–D221. doi:[10.1093/nar/gkt1128](https://doi.org/10.1093/nar/gkt1128)
110. Campbell MP, Packer NH (2016) UniCarbKB: new database features for integrating glycan structure abundance, compositional glycoproteomics data, and disease associations. *Biochim Biophys Acta*. doi:[10.1016/j.bbagen.2016.02.016](https://doi.org/10.1016/j.bbagen.2016.02.016)
111. Heinrichs S, Li C, Look AT (2010) SNP array analysis in hematologic malignancies: avoiding false discoveries. *Blood* 115(21):4157–4161. doi:[10.1182/blood-2009-11-203182](https://doi.org/10.1182/blood-2009-11-203182)
112. Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA (2008) Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutat Res* 659(1–2):147–157. doi:[10.1016/j.mrrev.2008.05.001](https://doi.org/10.1016/j.mrrev.2008.05.001)
113. Mi H, Thomas P (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol* 563:123–140. doi:[10.1007/978-1-60761-175-2_7](https://doi.org/10.1007/978-1-60761-175-2_7)
114. Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8(8):1551–1566. doi:[10.1038/nprot.2013.092](https://doi.org/10.1038/nprot.2013.092)
115. Mahmood AS, Wu TJ, Mazumder R, Vijay-Shanker K (2016) DiMeX: a text mining system for mutation-disease association extraction. *PLoS One* 11(4):e0152725. doi:[10.1371/journal.pone.0152725](https://doi.org/10.1371/journal.pone.0152725) PONE-D-15-16733 [pii]
117. Dingerdissen H, Motwani M, Karagiannis K, Simonyan V, Mazumder R (2013) Proteome-wide analysis of nonsynonymous single-nucleotide variations in active sites of human proteins. *FEBS J* 280(6):1542–1562. doi:[10.1111/febs.12155](https://doi.org/10.1111/febs.12155)
118. Swiss_Institute_of_Bioinformatics_Members (2016) The SIB swiss institute of bioinformatics' resources: focus on curated databases. *Nucleic Acids Res* 44(D1):D27–D37. doi:[10.1093/nar/gkv1310](https://doi.org/10.1093/nar/gkv1310)
119. Wan Q, Dingerdissen H, Fan Y, Gulzar N, Pan Y, Wu TJ, Yan C, Zhang H, Mazumder R (2015) BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database (Oxford)* 2015. doi:[10.1093/database/bav019](https://doi.org/10.1093/database/bav019)
120. Carithers LJ, Moore HM (2015) The Genotype-Tissue Expression (GTEx) project. *Biopreserv Biobank* 13(5):307–308. doi:[10.1089/bio.2015.29031.hmm](https://doi.org/10.1089/bio.2015.29031.hmm)
121. Hattori M (2005) Finishing the euchromatic sequence of the human genome. *Tanpakushitsu Kakusan Koso* 50(2):162–168
122. Jensen ON (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol* 8(1):33–41. doi:[10.1016/j.cbpa.2003.12.009](https://doi.org/10.1016/j.cbpa.2003.12.009)
123. Lu CT, Huang KY, Su MG, Lee TY, Bretana NA, Chang WC, Chen YJ, Chen YJ, Huang HD (2013) DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res* 41(Database issue):D295–D305. doi:[10.1093/nar/gks1229](https://doi.org/10.1093/nar/gks1229)
124. Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW, Mooney SD (2008) Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* 24(16):i241–i247. doi:[10.1093/bioinformatics/btn267](https://doi.org/10.1093/bioinformatics/btn267)
125. Lim YP (2005) Mining the tumor phosphoproteome for cancer markers. *Clin Cancer Res* 11(9):3163–3169. doi:[10.1158/1078-0432.CCR-04-2243](https://doi.org/10.1158/1078-0432.CCR-04-2243)
126. Yang JD, Roberts LR (2010) Hepatocellular carcinoma: a global view. *Nat Rev Gastroenterol Hepatol* 7(8):448–458. doi:[10.1038/nrgastro.2010.100](https://doi.org/10.1038/nrgastro.2010.100)
127. Norton PA, Comunale MA, Krakover J, Rodemich L, Pirog N, D'Amelio A, Philip R, Mehta AS, Block TM (2008) N-linked glycosylation of the liver cancer biomarker GP73.

- J Cell Biochem 104(1):136–149. doi:[10.1002/jcb.21610](https://doi.org/10.1002/jcb.21610)
128. Nakagawa T, Uozumi N, Nakano M, Mizuno-Horikawa Y, Okuyama N, Taguchi T, Gu J, Kondo A, Taniguchi N, Miyoshi E (2006) Fucosylation of N-glycans regulates the secretion of hepatic glycoproteins into bile ducts. *J Biol Chem* 281(40):29797–29806. doi:[10.1074/jbc.M605697200](https://doi.org/10.1074/jbc.M605697200)
 129. Marrero JA, Romano PR, Nikolaeva O, Steel L, Mehta A, Fimmel CJ, Comunale MA, D'Amelio A, Lok AS, Block TM (2005) GP73, a resident Golgi glycoprotein, is a novel serum marker for hepatocellular carcinoma. *J Hepatol* 43(6):1007–1012. doi:[10.1016/j.jhep.2005.05.028](https://doi.org/10.1016/j.jhep.2005.05.028)
 130. Gill DJ, Clausen H, Bard F (2011) Location, location, location: new insights into O-GalNAc protein glycosylation. *Trends Cell Biol* 21(3):149–158. doi:[10.1016/j.tcb.2010.11.004](https://doi.org/10.1016/j.tcb.2010.11.004)
 131. Schjoldager KT, Clausen H (2012) Site-specific protein O-glycosylation modulates proprotein processing—deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochim Biophys Acta* 1820(12):2079–2094. doi:[10.1016/j.bbagen.2012.09.014](https://doi.org/10.1016/j.bbagen.2012.09.014)
 132. Slawson C, Hart GW (2011) O-GlcNAc signalling: implications for cancer cell biology. *Nat Rev Cancer* 11(9):678–684. doi:[10.1038/nrc3114](https://doi.org/10.1038/nrc3114)
 133. Hart GW, Slawson C, Ramirez-Correa G, Lagerlof O (2011) Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annu Rev Biochem* 80:825–858. doi:[10.1146/annurev-biochem-060608-102511](https://doi.org/10.1146/annurev-biochem-060608-102511)
 134. Ohta T, Fukuda M (2004) Ubiquitin and breast cancer. *Oncogene* 23(11):2079–2088. doi:[10.1038/sj.onc.1207371](https://doi.org/10.1038/sj.onc.1207371)
 135. Kempen GI, Suurmeijer TP (1989) Depressive symptoms, invalidity and the use of professional home care by the elderly; replication and variations. *Tijdschr Gerontol Geriatr* 20(1):13–17
 136. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6–21. doi:[10.1101/gad.947102](https://doi.org/10.1101/gad.947102)
 137. Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3(6):415–428. doi:[10.1038/nrg816](https://doi.org/10.1038/nrg816)
 138. Baylin SB (2005) DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol* 2(Suppl 1):S4–11. doi:[10.1038/ncponc0354](https://doi.org/10.1038/ncponc0354)
 139. Herman JG, Latif F, Weng Y, Lerman MI, Zbar B, Liu S, Samid D, Duan DS, Gnarr JR, Linehan WM et al (1994) Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci USA* 91(21):9700–9704
 140. Esteller M, Garcia-Foncillas J, Andion E, Goodman SN, OF H, Vanaclocha V, Baylin SB, Herman JG (2000) Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N Engl J Med* 343(19):1350–1354. doi:[10.1056/NEJM200011093431901](https://doi.org/10.1056/NEJM200011093431901)
 141. Ranzinger R, Aoki-Kinoshita KF, Campbell MP, Kawano S, Lutteke T, Okuda S, Shinmachi D, Shikanai T, Sawaki H, Toukach P, Matsubara M, Yamada I, Narimatsu H (2015) GlycoRDF: an ontology to standardize glycomics data in RDF. *Bioinformatics* 31(6):919–925. doi:[10.1093/bioinformatics/btu732](https://doi.org/10.1093/bioinformatics/btu732)
 142. Montecchi-Palazzi L, Beavis R, Binz PA, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS (2008) The PSI-MOD community standard for representation of protein modification data. *Nat Biotechnol* 26(8):864–866. doi:[10.1038/nbt0808-864](https://doi.org/10.1038/nbt0808-864)

Analysis of Cysteine Redox Post-Translational Modifications in Cell Biology and Drug Pharmacology

Revati Wani and Brion W. Murray

Abstract

Reversible cysteine oxidation is an emerging class of protein post-translational modification (PTM) that regulates catalytic activity, modulates conformation, impacts protein–protein interactions, and affects sub-cellular trafficking of numerous proteins. Redox PTMs encompass a broad array of cysteine oxidation reactions with different half-lives, topographies, and reactivities such as S-glutathionylation and sulfoxidation. Recent studies from our group underscore the lesser known effect of redox protein modifications on drug binding. To date, biological studies to understand mechanistic and functional aspects of redox regulation are technically challenging. A prominent issue is the lack of tools for labeling proteins oxidized to select chemotype/oxidant species in cells. Predictive computational tools and curated databases of oxidized proteins are facilitating structural and functional insights into regulation of the network of oxidized proteins or redox proteome. In this chapter, we discuss analytical platforms for studying protein oxidation, suggest computational tools currently available in the field to determine redox sensitive proteins, and begin to illuminate roles of cysteine redox PTMs in drug pharmacology.

Key words Cysteine, Redox states, S-glutathionylation, Sulfoxidation, Proteomics, ROS, Chemical probes, Bioinformatics

1 Introduction

1.1 ROS as a Second Messenger

Cysteine is a unique amino acid because its thiol side chain is capable of participating in many types of chemical reactions which provides unique functionality to the modified proteins. One class of reaction is with reactive oxygen species (ROS). ROS are a persistent undercurrent in cells and central to the biology of aerobic organisms. A multitude of studies over the last decade and half have radically altered our perception of ROS from being agents of cellular distress (causing indiscriminate biomolecular oxidation, DNA damage, and cell death) to tightly regulated second messengers that modulate elemental cellular functions [1, 2]. Regulated production of ROS occurs in cells through the action of pro-oxidant enzymes such as oxidases (e.g., NADPH oxidase, NOX; dual

oxidase, DUOX; cyclooxygenase, COX; lysyl oxidase, LOX), superoxide dismutase (SOD), and mitochondrial electron transport chain (ETC) [2–5]. In cells, a delicate balance between ROS production and counteracting anti-oxidant systems is crucial for defining if the elicited protein oxidation is a stress response or a signaling cue for mediating functions such as proliferation, activation, differentiation, and migration. ROS can perform targeted protein oxidation that modify select protein cysteine residues for mediating signal transduction, biological functions, and drug activities [6]. Exacerbated ROS levels often lead to irreversible protein modifications and subsequent proteolysis. Taken together, ROS production is a highly evolved process that is central to many biological functions.

1.2 Post-Translational Modification to Cysteine Residues

Cysteine residues are well-suited to carry out sophisticated redox transactions because the sulfur atom of its sulfhydryl group is capable of having multiple valence states (-2 to $+6$) [5]. The reactivity of the residue is defined, in part, by the pK_a of its sulfhydryl group which describes the equilibrium between thiol and thiolate forms. The pK_a of the sulfhydryl group of an isolated cysteine amino acid is 8.2 [7] which is similar to the typical “unperturbed” cysteine residue in a protein (i.e., surface exposed, $pK_a \sim 8.5$ [4]). Physicochemical properties of cysteine residues such as pK_a and hydrogen bonding capability are influenced by the charge states of vicinal residues, surface localization, and access to solvent channels. These characteristics create a wide range of pK_a values as low as 2.5 for catalytic residues of cysteine proteases and protein phosphatases [1]. Oxidation of redox-sensitive cysteine residues in cells creates a range of modifications which can regulate signal transduction and biological functions. ROS can directly oxidize sensitive cysteine residues to yield distinct protein oxoacids (sulfenic, sulfinic, and sulfonic acids) [2, 4, 5, 8]. The tripeptide glutathione (γ -glutamyl-cysteinylglycine, GSH) is a highly abundant non-protein thiol constituting the cellular redox buffer. GSH can directly react with the sulfenyl oxoform of cysteine (P-SOH) to form a mix disulfide adduct (PSSG) [9]. Perturbations in ROS levels alter the balance between reduced (GSH) and oxidized (GSSG) forms of glutathione to create additional ways to form glutathione adducts. As such, cysteine residues are capable of being modified to many distinct chemical entities.

Although human proteome encodes 214,000 cysteine residues with 80 % of them predicted to have specific functional roles, our knowledge of targeted protein oxidation is still quite inceptive [2]. The analysis of cysteine redox biology is further complicated by the presence of multiple redox-sensitive cysteine residues with different reactivities in a protein that can make it challenging to measure functional effects of post-translational oxidation induced by specific cysteine residues. To convincingly characterize redox modifications as

Table 1
Redox modifications of protein cysteine residues and corresponding labeling methods

Protein cysteine modification	Functional group	Probes/technique (ref)
S-Glutathionylation	Cys-S-S-G	Biotin switch, PSSG antibodies, MS analysis [10, 11]
S-Cysteinylation	Cys-S-S-Cys	Biotin switch, MS analysis [10, 11]
S-Sulphydration	Cys-S-SH	Modified biotin switch, MS analysis [12]
Sulfenic oxidation	Cys-S-OH	Dimedone/derivatives (DCP-Bio1, DAz1; OxICAT), MS analysis [8, 13, 14]
Sulfinic oxidation	Cys-SO ₂ H	MS analysis [15]
Sulfonic oxidation	Cys-SO ₃ H	MS analysis, high-affinity polyarginine-coated nanodiamonds [16]

physiologically relevant, a combination of complimentary approaches is necessary (i.e., biochemical, cellular, and computational). Molecular biology techniques such as site-directed mutagenesis can help validate roles of individual protein cysteine residues and are often combined with other approaches for functional assessment. A list of commonly known covalent and non-covalent cysteine redox modifications and currently known tracers for their labeling is assembled in Table 1. In this chapter, we summarize common analytical approaches taken to label and identify oxidized protein targets in vitro and briefly touch upon emerging bioinformatics tools for proteomics analysis of oxidatively modified proteins as techniques and tools for investigating targeted protein redox modifications are still emerging. Since the biochemical characteristics of oxidized proteins can vary significantly depending on the number, location, pK_a and other molecular properties of redox active cysteine residues, we describe general experimental schemes that can be employed to identify and validate protein oxidation. As cysteine residues can be highly nucleophilic, other reactive capabilities occur but are outside the scope of this chapter and reviewed elsewhere [2, 17].

2 Materials

1. Biochemical and Cellular Materials

Hydrogen peroxide, H₂O₂; oxidized glutathione, GSSG; 3-amino-1,2,4-triazole, 3-AT; reduced glutathione, GSH; L-buthionine sulfoximine, BSO; VAS-2870; PEG-catalase; Diamide; rotenone; *N*-acetyl cysteine, NAC; low passage cell lines; cell culture medium; 0.25 % Trypsin; 1X cell lysis buffer – RIPA/Tris chloride; HEPES; highly purified recombinant proteins; DTT; TCEP; DMSO; anti-glutathione [D8]

monoclonal antibody; anti-biotin antibody; anti-avidin antibody; Protein A-sepharose 4B; iodoacetamide, IAA; *N*-ethyl maleimide, NEM; S-methyl methanethiosulfonate, MMTS; ascorbate; biotinylated IAA; 6-(iodoacetamido)fluorescein; S-glutathionylated protein detection kit; SILAC amino acids—l-lysine-¹³C₆,¹⁵N₂ hydrochloride, L-arginine-¹³C₆,¹⁵N₄ hydrochloride, l-lysine-¹³C₆ hydrochloride, l-arginine-¹⁵N₄ hydrochloride, l-lysine hydrochloride, l-arginine hydrochloride; dialyzed serum; penicillin-streptomycin; corresponding cell culture medium deficient in lysine and arginine DMEM/RPMI; 2',7'-dichlorofluorescein, H₂DCFDA/DCF; epidermal growth factor, EGF; 0.22 μm filter; platelet-derived growth factor BB, PDGF; protein and peptide identification programs.

2. Protein ID (UniProt ID#, NCBI Accession #); Internet access to

PROPKA 3.0: <http://propka.org/>

CMD: <http://birg4.fbb.utm.my/cmd/>

UniProt: www.uniprot.org

Protein Data Bank (PDB): <http://www.rcsb.org/pdb/>

RedoxDB: <http://biocomputer.bio.cuhk.edu.hk/RedoxDB/>

dbGSH: <http://csb.cse.yzu.edu.tw/dbGSH>

DiANNA 1.1: <http://clavius.bc.edu/~clotelab/DiANNA/>

DISULFIND: <http://disulfind.dsi.unifi.it/>

DBCP: <http://biomedical.ctust.edu.tw/edbcp/>

CYSPRED: http://gpcr.biocomp.unibo.it/cgi/predictors/cyspred/pred_cyspredcgi.cgi.

3 Methods

3.1 Identifying Redox-Sensitive Protein Cysteine Residues

As discussed above, the malleability of the cysteine thiol group enables it to assume multiple oxidation states with different chemical reactivities and stabilities [5]. Combined qualitative and quantitative approaches are required to understand contribution of select oxidation events to both protein and cellular functions at large. Chemical biology approaches to investigate redox signaling include in vitro chemical modification of proteins, cellular treatments with redox inhibitors, and labeling of select protein cysteine oxoforms with specific chemical tags for subsequent quantitative or qualitative redox analysis. Described below are protocols for (1) post-translationally modifying proteins with oxidants in vitro, (2) perturbing redox levels in cells with different inhibitors, and (3) labeling redox-sensitive cysteine residues in proteins.

3.1.1 Analysis of Redox Modification to Isolated Proteins

Purified recombinant proteins (or proteins immunoprecipitated from cells) can be treated *in vitro* with oxidized glutathione, GSSG to S-glutathionylate [18, 19] or with hydrogen peroxide, H₂O₂, to oxidize (e.g., -SOH, -SO₂H, -SO₃H, -S-S) [13, 20, 21] susceptible protein cysteine residues (*see Note 1*). Described below are protocols (in two parts) for select redox modifications of recombinant purified proteins and biotin switch assay.

Analysis of Isolated Proteins by Mass Spectrometry or Western Blot

1. *In vitro* redox modification reactions can be carried out under alkaline conditions (pH >7) at a molar ratio of 1:250 (protein:GSSG) or 1:20 (protein:H₂O₂) in corresponding buffers (e.g., RIPA, Tris, HEPES). Briefly, 1 μM recombinant protein can be modified in 50–100 μL reaction volumes under alkaline conditions using 250 μM GSSG (reaction volumes can be scaled up proportionately) (*see Note 1*).
2. The protein-oxidant reaction is incubated for variable amounts of time depending on the redox sensitivity of protein cysteine(s), extent of modification desired, protein stability at room temperature, and intended application of the modified protein. Shorter incubations of 2–4 h can be conducted at room temperature while longer incubations can be conducted at 4 °C. In general, the molar ratio of protein:oxidant, incubation time, and temperature for individual proteins can vary significantly depending on their redox sensitivity and always should be optimized.
3. Eliminating excess reagents (oxidants) from the reaction and/or termination to minimize adventitious oxidation can be accomplished by passing proteins through appropriate desalting spin columns that are pre-equilibrated with identical buffer but at a neutral pH to render cysteine residues less nucleophilic. For instance, a 100 μL reaction of protein and GSSG (1:250 molar ratio) in pH 8.5 HEPES can be desalted through 1 mL spin columns that are pre-equilibrated with pH 7.0 HEPES buffer.
4. Protein S-modifications are confirmed by multiple methods. Intact mass analysis of the modified proteins by mass spectrometry (ESI/MALDI) is a direct method of detection. Expected mass shifts for modified cysteine residues can be noted from the deconvoluted spectra. For instance, +305 shift per glutathionylated cysteine (-SSG) adduct, +119 shift per cysteinylated adduct, +16 mass shift per sulfenylated (-SOH) adduct, +32 mass shift per sulfinylated (-SO₂H), adduct and +48 mass shift per sulfonylated (-SO₃H) adduct, etc. Further characterization of the modified cysteine residues can be achieved by MS/MS analysis wherein the S-oxidized protein can be enzymatically digested and the peptides containing ions for the respective S-modifications are searched.

- An alternative way of probing for protein S-glutathionylation is through western blotting of the modified proteins using anti-glutathione [D8] monoclonal antibodies (Virogen, Abcam) that specifically recognize glutathione-conjugated proteins under non-reducing conditions.

Using multiple approaches to validate S-oxidation of target proteins is essential to eliminate artifacts. Developing cellular equivalents for in vitro redox-modified proteins is equally important to delineate biological significance of the redox-reactive cysteine residues.

Biotin Switch Method

The biotin switch technique was first introduced to identify S-nitrosylated cysteine residues and has three steps: (1) blocking unmodified cysteine residues, (2) converting redox-modified cysteine residues to their thiol form, and (3) labeling unmasked cysteine residues (Fig. 1) [10]. Briefly, cysteine residues with free thiols are blocked with an alkylating reagent (e.g., IAA, NEM) followed by treatment of the redox-modified cysteine residues with a reducing agent (ascorbate, DTT, TCEP). The reduced cysteine residues (formerly redox-modified) are labeled with either a modified derivative of the initial alkylating agent (biotinylated IAA, NEM, or fluorochrome-conjugated alkylating agents such as fluorescein IAA) or an altogether different alkylating agent. This procedure is now routinely applied to detect other redox cysteine modifications such as S-glutathionylation, S-nitrosylation, sulfenylation followed by immunoaffinity capture and combined with mass spectrometry-based protocols for cysteine identification, immunofluorescence imaging, and flow cytometry applications. (*see Note 2*).

In continuation of the previous protocol,

- 50 μM S-oxidized recombinant proteins can be incubated with an alkylating agent to block unmodified S-thiols (40–50 mM NEM for 1–2 h at RT in dark) followed by reduction of S-modified cysteine residues with TCEP or DTT (tenfold molar excess, i.e., 500 μM at RT for 1 h in dark).
- Re-alkylation can be conducted with either fluorescein IAA/ biotinylated IAA (Life Technologies) or a different alkylating agent than used in the first alkylation step (IAA or MMTS for 1–2 h at RT in dark).
- Excess reagents are removed by passing the protein/reaction mixture through desalting spin columns and the originally S-oxidized cysteine residues can be detected by MS/MS analysis.

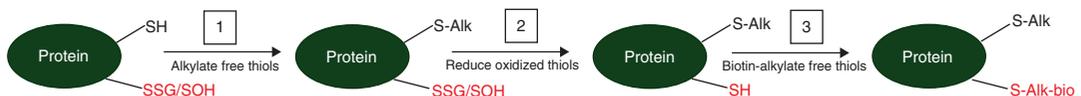


Fig. 1 Overview of the biotin switch method

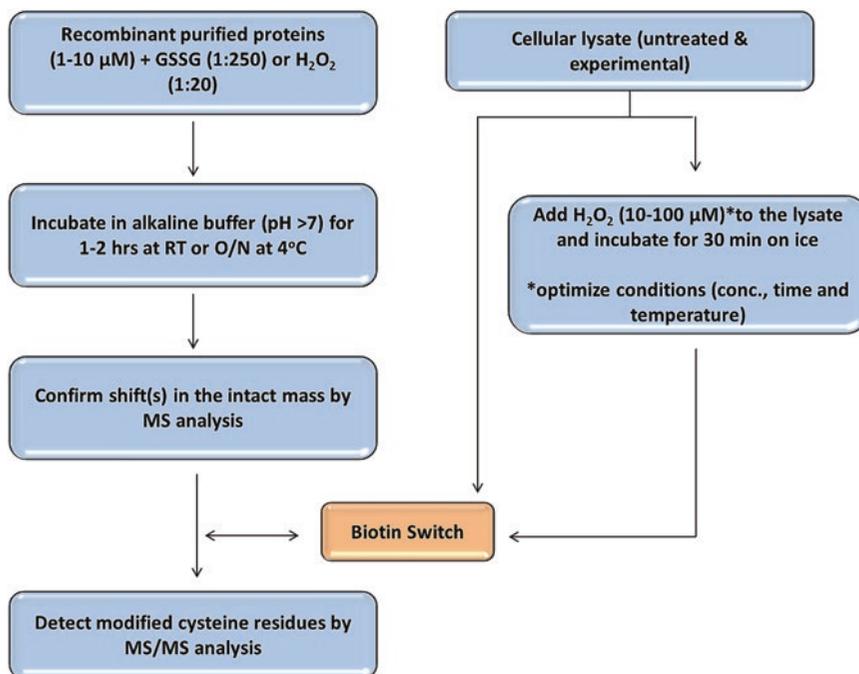


Fig. 2 The in vitro protein S-modification workflow

Biotin switch can be performed on S-modified proteins prior to MS/MS analysis. Similar strategy can be used for modifying cellular proteins [11], however the concentration of oxidants (GSSG, H_2O_2), time of incubation and temperature for modifying cellular lysates can be vastly different than for in vitro protein modifications. The in vitro protein S-modification workflow is summarized in Fig. 2.

3.1.2 Chemical Biology Analysis in Cells with Redox Reagents

Common technical approaches for perturbing cellular redox environment include treatments with small molecule pro-oxidants such as hydrogen peroxide (H_2O_2), catalase inhibitor (3-AT), GSH biosynthesis inhibitor L-buthionine sulfoximine (BSO); small molecule antioxidants such as NOX inhibitors (VAS-2870), ascorbate, and enzymes such as PEG-catalase. Effective concentrations and incubation times for some of the redox inhibitors based on prior literature [13, 15] are as follows: catalase inhibitor 3-AT at 10–15 mM for 16–20 h incubation at 37 °C; H_2O_2 treatment at 0.1–1 mM for 20–30 min at room temperature or 37 °C; Diamide at 25–50 μM for 30 min at room temperature or 37 °C; BSO at 0.5–20 mM for 16–20 h incubation at 37 °C; VAS-2870 at 5–20 μM for 16–20 h incubation at 37 °C; the PEGylated form of antioxidant enzyme catalase, i.e., PEG-catalase at 250–1000 U/mL for 2–5 h incubation at 37 °C; mitochondrial electron transport chain (ETC) inhibitor Rotenone at 1–20 μM for 16–20 h incubation at 37 °C and the antioxidant *N*-acetyl cysteine at 1–20 mM for 16–20 h incubation

at 37 °C. Always adjust the pH of cell culture media after resuspending the reagents since the pH of the medium will turn acidic with certain inhibitors such as BSO and NAC. Described below is a general method for different redox treatments and the concentrations of individual reagents should be optimized to specific cell lines and to the experimental set up (*see Note 3*).

Chemical Biology
Approaches to Investigate
Redox Biology

1. Culture cells to desired confluence in the corresponding complete cell culture medium. As a general guideline, 60–70 % confluent cells can be used for adding inhibitors for overnight treatments (ensure that the cells do not get fully confluent prior to inhibitor treatments).
2. Freshly prepare cell culture medium with the required redox inhibitors as mentioned above (complete or serum-free medium can be used depending on the need for synchronization).
3. Additional treatments with mitogens, cytokines, or other inhibitors should be factored in for individual protocols. For instance, growth factor stimulations (EGF, PDGF at 20–100 ng/mL) typically involve shorter time points (minutes) and can be performed after redox treatments on the following day in fresh serum-free culture medium. Combination of redox inhibitors with other cellular inhibitors should be tested for individual cell lines.
4. Effects of these redox treatments on cellular ROS can be tested using the standard, commercially available fluorescence-based DCF reagent (Life Technologies) by spectrophotometric, imaging, or flow cytometry-based read outs.

Inhibitor-treated cells from above can be subsequently (1) lysed in respective cell lysis buffers (with and/or without chemical tags for S-oxidized cysteine residues) for western analysis, immunoprecipitation studies, (2) immunofluorescence-imaged for respective redox PTM such as S-glutathionylation (with co-localization markers), or (3) sorted via flow cytometer for enriching in the labeled cell populations as described in Fig. 3.

3.1.3 Labeling Oxidized
Protein Cysteine Residues
in Cells

Chemically tagging S-modified residues can facilitate more reliable detection and quantitation of oxidized proteins. The tools described here can be coupled with different enrichment strategies at different stages of the protocol to create unique workflows for specific applications and thus make the protocols appreciably resilient. We discuss common technical workflows with a focus on sulfenylated and S-glutathionylated proteins.

Identification of Proteins
with Cysteine PTMs
in Cells

1. Cells cultured in 100 or 150 mm (or other relevant size) tissue culture plates can be lysed in 0.5–1 mL corresponding 1× cell lysis buffer (50–100 mM RIPA, HEPES, Tris, etc.) with protease and phosphatase inhibitors along with low concentration

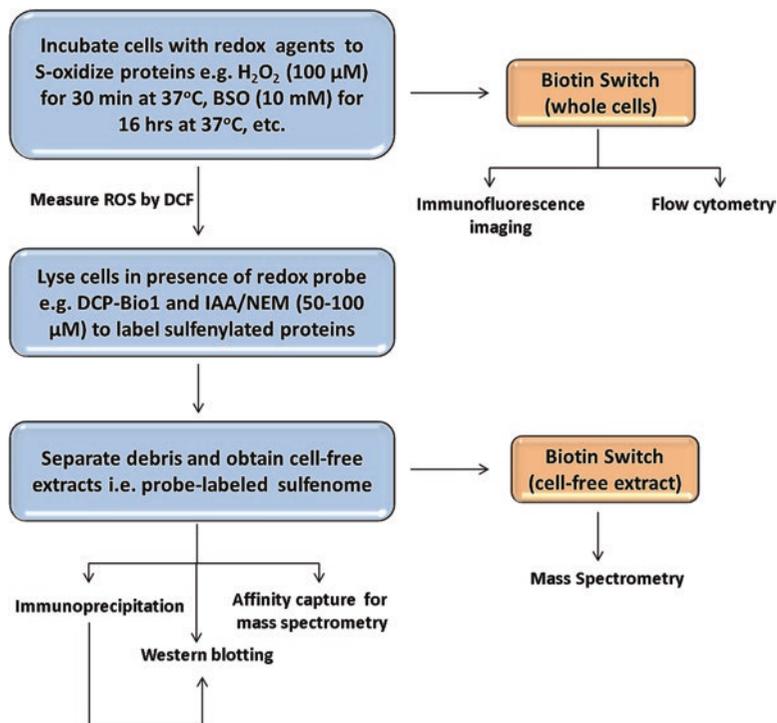


Fig. 3 The workflow of labeling oxidized proteins in cells

of alkylating agents (IAA or NEM, 10–50 μM) to prevent artifactual oxidation, and/or, ii) chemical probes for rapidly labeling S-modified proteins such as DCP-Bio1 [14] for sulfenylated residues (0.1–1 mM).

- The lysates should be incubated for 30–60 min on ice and spun down ($10,000 \times g$ for 10 min at 4°C) to remove debris and obtain clear cell-free extracts which could be used directly for biotin switch as described in Subheading 3.1.2 or for enrichment/affinity capture of sulfenylated proteins with antibodies against the tag (biotin/avidin) for other applications (western blotting, immunoprecipitation, MS studies).
- Detection of the labeled cysteine residues can be achieved directly by immunofluorescence (IF) imaging with fluorophore-based alkylating agents and indirectly by fluorophore conjugated biotin or streptavidin antibodies, two-dimensional difference gel electrophoresis (DIGE) and/or mass spectrometry. Commercially available kits enable easier labeling of S-glutathionylated proteins in cells and in situ and detection by IF imaging, immunocapture for western blotting, and MS analysis.

A revised approach for trapping oxidized proteins while minimizing artifactual protein oxidation termed PROP was recently developed for redox proteomics studies. The PROP technique

utilizes acid precipitation strategy to rapidly fix cells for more efficient detection of oxidized proteins [22]. Multiple reagents and methods have emerged over the last decade that enable selective labeling of redox-modified cysteine residues in proteins (Table 1). The workflow of labeling oxidized proteins in cells using these probes is described in Fig. 3.

3.2 Redox Proteomics Analysis

Getting a global snapshot of the redox-modified proteome, i.e., “redoxome” is critical to understand and fully appreciate the pervasiveness of this PTM in a variety of contexts. Proteomics analysis has evolved from being qualitative [23, 24] to now a highly quantitative tool [25–27] that enables fine dissection of disease biology [28, 29]. For quantitative assessment of the redox proteome, stable isotope labeling by amino acids in cell culture (SILAC) can be performed prior to proteomics studies [3, 30]. Detailed protocols for labeling cells with SILAC and subsequent MS/MS analysis strategies are detailed by Mann *et al.* [31]. Here, we focus on the application of SILAC-labeled cells for redox proteomics studies in context of the workflow described in Fig. 4 (*see Note 4*).

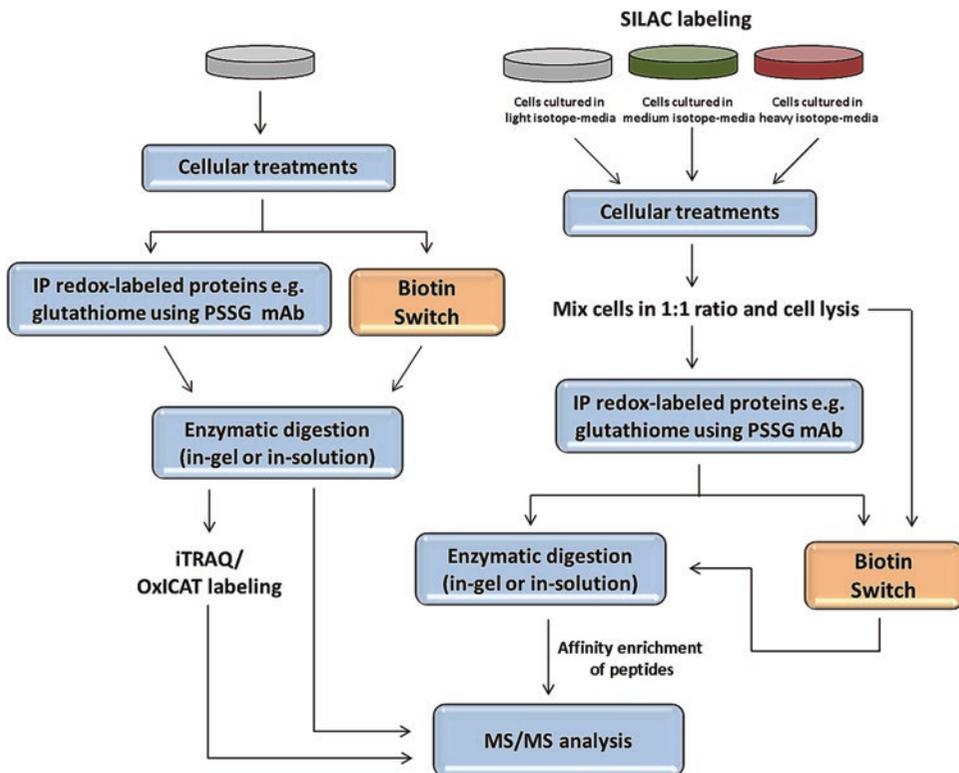


Fig. 4 A simplified workflow for redox proteomics studies

3.2.1 *Quantitation
of the Cellular Cysteine
Redox Proteome by SILAC*

1. Prepare SILAC growth medium by separately adding from the filter-sterilized stocks (1000× stocks) of light, medium, and heavy isotopes of lysine and arginine to the appropriate cell culture medium (DMEM/RPMI) which is deficient in the two amino acids followed by addition of dialyzed FBS (10 %) and penicillin-streptomycin (1 %).
2. Establish the cell line of interest (low passage number strongly recommended) in culture and subculture it 1–2 times (maximum). Once confluent, split cells equally into three separate SILAC media (light, medium, and heavy isotope supplemented) and treat them as separate cell lines this point forward.
3. Subculture the cell lines every 2–3 days or at appropriate intervals (70–80 % confluence) in respective isotopically supplemented medium for a minimum of 5 passages to ensure complete incorporation of amino acids in cellular proteome. More detailed SILAC labeling and verification strategy is described in protocols from Mann's laboratory [31].
4. Conduct cellular experiments/treatments (growth factors, cytokines, oxidants, small molecules, etc.) in the labeled population of cells. For instance, if the objective of the study is to understand effects of glutathione depletion and peroxide treatments on cellular glutathione/sulfenome, then set up cell culture for three treatment groups—untreated control/vehicle group, BSO treatment group and H₂O₂ treatment group in the three SILAC supplemented media conditions—light, medium, and heavy isotope respectively.
5. Add BSO and H₂O₂ at pre-optimized concentration and incubation time (as discussed in Subheading 3.1.2, step #2) respectively to subconfluent cells that are cultured in medium isotope and heavy isotope supplemented SILAC media while maintaining the vehicle group in light isotope containing medium.
6. At the end of respective treatments, wash cells twice with 1× PBS, trypsinize and count. Cells from each condition can be mixed in 1:1 ratio and then lysed in appropriate lysis buffers (1× RIPA, 1× HEPES, etc.) containing Cys-SOH labeling probe such as DCP-Bio1 (Subheading 3.1.3, step #1) for sulfenome analysis. Process lysates to obtain cell-free extracts as discussed in Subheading 3.1.3, step #2) followed by estimation of protein concentration. Alternatively, the cells can be separately lysed first and the extracts can be mixed in 1:1 ratio (will depend on the nature of experiment).
7. To isolate the respective redox-modified subproteomes, perform affinity enrichment. Briefly, to the clear extracts add anti-glutathione monoclonal antibody (PSSG) at optimized ratio of protein to antibody (e.g., 100 µg protein extract can be mixed with 5–10 µg PSSG antibody) for enriching in glutathionylated

proteins or add anti-biotin/streptavidin antibody at optimized ratio for capturing DCP-Bio1-tagged sulfenylated proteins. Incubate the protein-antibody mixture overnight on a nutator (low spin setting) at 4 °C.

8. Next day, add protein A sepharose or agarose beads to the respective pull downs (at pre-optimized enrichment conditions such as incubation time and temperature with the beads). Following incubation, wash the beads carefully but stringently in the corresponding 1X lysis buffer for at least three times to eliminate nonspecific protein interactions and obtain cleaner preps. Gel loading tips can be used for aspirating lysis buffers during wash steps to avoid losing protein-bound beads.
9. Depending on the nature of isolated subproteome and the experimental objective, the protein-bound beads can be (1) enzymatically digested for MS/MS studies; (2) subjected to biotin switch prior to enzymatic digestion, i.e., reduction and alkylation steps as described in Subheading 3.1.1; (3) eluted from the beads by boiling in sample buffer directly or after performing biotin switch.
10. The proteins/eluates collected after cooling and spinning down the boiled samples can be run on SDS-PAGE gels to separate the proteins. The protein gel can be subsequently stained with MS-compatible protein stains (Coomassie brilliant blue, silver stain, etc.) for few hours at RT followed by destaining until the stained protein bands are visible on a clear background.
11. Excise each protein band from the gel carefully without mixing with protein bands above or below from each sample lane and collect them in well-numbered protein LoBind microfuge tubes. For error minimization, the stained gel could be imaged first and individual bands can be numbered and excised accordingly.
12. The excised gel pieces can be enzymatically digested with trypsin (or other relevant protease) following standard digestion and processing protocols [31].
13. The extracted, dried peptides (containing redox-modified cysteine residues) are then processed for LC and MS/MS analysis. Search the peptides using the mass spectrometer-specific database and factor-in the mass difference between light, medium, and heavy isotope containing peptides for deducing the number of SILAC amino acids present in the specific peptide. Quantitation of the three peptides can be achieved by noting the signal intensity and acquisition time for individual isotope-peptide and performing ratio analysis for all measurable peptides for a select protein to quantitate its abundance in the three treatment groups.

In addition to this metabolic labeling/SILAC strategy, chemical labeling approaches such as iTRAQ (isobaric tags for relative and absolute quantitation) and ICAT/OxICAT also facilitate

quantitation in proteomics studies especially of primary tissues that cannot be metabolically labeled. iTRAQ is a highly multiplexed technique (8-plex) that involves labeling of enzymatically digested peptides with isobaric reagents that label primary amines of peptides which can be selectively quantitated [32]. Each isobaric label has a peptide reactive group, a neutral balance group to maintain overall mass (145 Da) and a unique reporter group that gets released upon MS/MS fragmentation to yield ions of distinct m/z . Another proteomics approach termed ICAT (isotope coded affinity tag) involves an isotopically coded (^{12}C and ^{13}C), thiol-reactive tag (iodoacetamide) linked to a cleavable biotin probe. A variant of this approach termed OxICAT (oxidation ICAT) offers information on extent of protein oxidation by utilizing the light and heavy carbon tags sequentially in a biotin switch set up, i.e., using ^{12}C for first round of thiol blocking and ^{13}C for post-reduction alkylation (biotin switch format) of oxidized cysteine residues [33]. Proteomics workflow for analyzing redox PTMs qualitatively and quantitatively incorporating biotin switch has been summarized well by Van Eyk [15]. A simplified workflow for redox proteomics studies is shown in Fig. 4.

3.3 Computational Tools for Predicting Protein Oxidation

A brief discussion of the advances and constraints in developing computational tools for characterizing functions of redox sensitive cysteine residues is presented in this chapter but the topic is comprehensively treated elsewhere [34]. Based on the known biochemical functions, cysteine residues are categorized into four groups—metal binding, catalytic, regulatory, and structural [34]. Computational approaches to predict redox reactive cysteine residues utilize structural and physico-chemical information derived from the knowledge of previously characterized catalytic cysteine residues. However, challenges arise when performing predictive analyses for non-catalytic cysteine residues as well as for those residues that fall into multiple functional classes. Further, the dynamic cellular environment exposes protein cysteine residues to different redox compartments while trafficking which can alter nucleophilicity of sulfhydryl group(s) to affect their reactivity within the sub-cellular niche adding another complexity to development of bioinformatics tools for realistic prediction of a protein's redox sensitivity [5]. Given the intrinsic broad array of cysteine oxoforms (covalent and non-covalent), several considerations are required for developing methods for more reliable and predictive readouts. Bioinformatics tools for predicting cysteine oxidation states and disulfide associations are still emerging and the field is nascent [35]. As such, a rudimentary approach integrating known computational tools in a step-by-step format for predictive analysis of redox active cysteine residues is suggested below.

3.3.1 Workflow for Bioinformatic Analysis of Redox Cysteine Biology

1. To determine redox sensitivity of a protein, one of the first steps could involve estimating the $\text{p}K_a$ of cysteine residues by a tool such as PROPKA that quantitates $\text{p}K_a$ values of ionizable

residues [36]. Since some analyses consider cysteine residues with depressed pK_a values as a direct indication of the protein's redox sensitivity, residues with lower pK_a can be presumed more reactive [1, 26].

- To identify structural and sequence elements/motifs around redox-sensitive cysteine residues, tools such as CMD (cysteine motif database) can be used. Cysteine motif database (CMD) includes a compilation of cysteine residues and the motifs associated with their secondary structures [37]. Utilizing protein data from a protein sequence and function database (e.g., Uniprot) and the protein data bank (PDB), information about free and disulfide bonded cysteine residues, frequency of their occurrence and coefficient of disulfide bonding and motif sequences (about 3 million) are compiled. For example, the Uniprot ID for the phosphatase PTEN can be obtained from the Uniprot database as shown in Fig. 5.

PDB ID for PTEN can be searched on the PDB site as follows (Fig. 6). The search displays available structures for all PTEN proteins. 5BZZ is the PDB ID for the reduced PTEN protein that can be used for subsequent applications.

- To determine oxidation state(s) of redox sensitive cysteine residues, data mining tools such as RedoxDB, dbGSH can be used. RedoxDB, a curated database of experimentally validated redox-modified proteins has a compilation of around 2200 cysteine residues reported modified from 1200 known oxidized proteins from over 90 organisms [36]. Another database dbGSH lists over 200S-glutathionylated cysteine residues from over 100 glutathionylated proteins as derived by literature mining [38]. This database also provides information about solvent accessibility, secondary and tertiary structures, gene

UniProtKB results

Filter by: Reviewed (158) items, Unreviewed (2,763) items

Popular organisms: Human (101), Mouse (58), Rat (33), Zebrafish (27), Rice (26), Other organisms

Search terms: Filter "pten" as: disease (1), gene name (173)

Entry	Entry name	Protein names	Gene names	Organism	Length	
<input checked="" type="checkbox"/>	P00484	PTEN_HUMAN	Phosphatidylinositol 3,4,5-trisphosphatase	PTEN, PPMAC1, TEP1	Homo sapiens (Human)	403
<input type="checkbox"/>	O08586	PTEN_MOUSE	Phosphatidylinositol 3,4,5-trisphosphatase	Pten, Ptnac1	Mus musculus (Mouse)	403
<input type="checkbox"/>	Q9Y0B6	Q9Y0B6_DROME	PTEN1	Pten, PTEN, CG5671, Dmel_CG5671	Drosophila melanogaster (Fruit fly)	418
<input type="checkbox"/>	Q9V3L4	Q9V3L4_DROME	PTEN3	Pten, PTEN, CG5671, Dmel_CG5671	Drosophila melanogaster (Fruit fly)	509
<input type="checkbox"/>	O54857	O54857_RAT	Phosphatase and tensin homolog, isoform 1	Pten, PTEN/PPMAC1, rCG_47874	Rattus norvegicus (Rat)	403
<input type="checkbox"/>	Q9Y0B5	Q9Y0B5_DROME	PTEN2	Pten, PTEN, CG5671	Drosophila melanogaster (Fruit fly)	511
<input type="checkbox"/>	Q7KMQ6	Q7KMQ6_DROME	IP16020p	Pten, CG5671, Dmel_CG5671	Drosophila melanogaster (Fruit fly)	514
<input type="checkbox"/>	Q8T957	PTEN_DICDI	Phosphatidylinositol 3,4,5-trisphosphatase	pten, ptenA, DOB_G0286557	Dictyostelium discoideum (Slime mold)	533
<input type="checkbox"/>	Q9BXN7	PINK1_HUMAN	Serine/threonine-protein kinase PINK1	PINK1	Homo sapiens (Human)	581
<input type="checkbox"/>	F1P8M8	F1P8M8_CANLF	Phosphatidylinositol 3,4,5-trisphosphatase	PTEN	Canis lupus familiaris (Dog) (Canis familiaris)	381

Fig. 5 UniProt search for the phosphatase PTEN

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

RCSB PDB An Information Portal to 118748 Biological Macromolecular Structures

PTEN Go

Advanced Search | Browse by Annotations | Search History (1) | Previous Results (18)

18 Structures 10 Unreleased Structures 9 Citations 12 Ligands

Search Parameter: Text Search for: pten Refine Search Save Search to MyPDB

Refinements

ORGANISM

- Homo sapiens (8)
- Mus musculus (4)
- Ciona intestinalis (3)
- Rattus norvegicus (2)
- Bos taurus (1)

UNIPROT MOLECULE NAME

- Phosphatidylinositol 3,4, ... (6)
- Voltage-sensor containing ... (3)
- Protein deglycase DJ-1 (2)
- Partitioning defective 3 ... (2)
- Phosphatidylglycerophosph ... (2)
- Phosphatidylinositol 4,5- ... (1)
- Protein tyrosine phosphat ... (1)
- Refine Query

TAXONOMY

- Eukaryota only (18)

EXPERIMENTAL METHOD

- X-ray (15)
- Solution NMR (3)

Currently showing 1 - 16 of 18 Page: 1 of 1 Displaying 25 Results

View: Detailed Reports: Select one... Sort: Sort by... Download Files

5BUG Download File View File

Crystal structure of human phosphatase PTEN oxidized by H₂O₂

[Lee, C.-U., Bier, D., Hennig, S., Grossmann, T.N.](#)

Redox Modulation of PTEN Phosphatase Activity by Hydrogen Peroxide and Bisperoxidovanadium Complexes. (2015) *Angew.Chem.Int.Ed.Engl.* 54: 13796-13800

Released: 2015-10-07 Method: X-RAY DIFFRACTION Macromolecule Content Phosphatidylinositol 3,4,5-tri ... (protein)

Resolution: 2.40 Å Unique Ligands: 1 TLA

Residue Count: 1256

5BZX Download File View File

Crystal structure of human phosphatase PTEN treated with a bisperoxovanadium complex

[Lee, C.-U., Bier, D., Hennig, S., Grossmann, T.N.](#)

Fig. 6 Protein Database (PDB) search for PTEN crystal structures

ontology among others to enable structural and functional analyses. For instance, to determine oxidation state of the cysteine residues in PTEN using RedoxDB, the PDB ID 5BZZ can be entered in the search term as below (Fig. 7).

RedoxDB suggests S-nitrosylation of Cys83 and an intramolecular disulfide between Cys71 and Cys124 residues in PTEN (Fig. 8).

dbGSH shows S-glutathionylation of the two residues Cys71 and Cys124 in PTEN that are predicted to form an intramolecular disulfide (Fig. 9).

- To predict cysteine residues involved in disulfide bond formations, prediction tools such as DiANNA, DISULFIND, DBCP, and CYPRED can be used. DiANNA is a database for predicting potential cysteine disulfides in proteins and the presence of free versus ligand-bound cysteine residues (metals such as Zn and Fe) [39]. DISULFIND predicts disulfide formation with a confidence value based on the protein sequence input [40]. DBCP is another tool that helps predict disulfide connectivity between cysteine residues for unknown proteins based on the protein sequence [41]. PTEN protein cysteine residues Cys71 and Cys124 are also predicted to form an intramolecular disulfide

RedoxDB: a curated database of protein oxidative modification

Home Search BLAST Browse Download Submit Help

Entry: PTEN_HUMAN

General information

Gene names	PTEN MMAC1 TEP1
Protein names	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN Mutated in multiple advanced cancers 1 Phosphatase and tensin homolog
Organism	Homo sapiens (Human)
Function	Tumor suppressor. Acts as a dual-specificity protein phosphatase, dephosphorylating tyrosine-, serine- and threonine- phosphorylated proteins. Also acts as a lipid phosphatase, removing the phosphate in the D3 position of the inositol ring from phosphatidylinositol 3,4,5-trisphosphate, phosphatidylinositol 3,4-diphosphate, phosphatidylinositol 3- phosphate and inositol 1,3,4,5-tetrakisphosphate with order of substrate preference in vitro PtdIns(3,4,5)P3 > PtdIns(3,4)P2 > PtdIns3P > Ins(1,3,4,5)P4. The lipid phosphatase activity is critical for its tumor suppressor function. Antagonizes the PI3K- AKT/PKB signaling pathway by dephosphorylating phosphoinositides and thereby modulating cell cycle progression and cell survival. The unphosphorylated form cooperates with AIP1 to suppress AKT1 activation. Dephosphorylates tyrosine-phosphorylated focal adhesion kinase and inhibits cell migration and integrin-mediated cell spreading and focal adhesion formation. Plays a role as a key modulator of the AKT-mTOR signaling pathway controlling the tempo of the process of newborn

Fig. 7 RedoxDB search for PTEN

Sequence

Length 403 AA Download [FASTA](#)

```
>PTEN_HUMAN 71:intra-disulfide; 83:S-nitrosylation; 124:intra-disulfide;
MTAIIKEIVSRNKRRYQEDGFDLDTYIYPNIIAMGFPAERLEGVYRNNI DDVVRFLDSK
HKNHYKIYNLC AERHYDTAKFNC RVAQYPFEDHNPQLELIKPFCELDLQWLSEDDNHVA
AHC KAGKGRGTGVMICAYLLHRGKFLKAQEALDFYGEVTRDRKKGVTIP SQRRYVYYSY
LLKNHLDYRVPALLFHKMMFETIPMFSGGTCNPQFVVCQLKVKIYSSNSGPTRRREDKFMY
FEFPQPLPVCGLKVEFFHKQNKMLKDKMFHWNTFFIPGPEETSEKVENGLCDQEI
DSICSIERADNDKEYLVLT LTRKNDLDKANKDKANRYFSPNFVKVLYFTKTVEEPSNPEAS
SSTSVPDPVSDNEPDHYRSDT L TSDSPENEPFDEDDQHTQITKV
```

Modification

Protein oxidative modification

Modification	Source of data	Reference
S-nitrosylation	case_study	1-4
intra-disulfide		

Cysteine residue modification

Position	Flanking Sequence	Modification	Source of data	Reference
71	HKNHYKIYNLC AERHYDTAKF	intra-disulfide	case_study	1, 2, 3
83	ERHYDTAKFNCRVAQYPFEDH	S-nitrosylation	case_study	4
124	EDDNHVAIIHCKAGKGRGTGVM	intra-disulfide	case_study	1, 2, 3

Fig. 8 Sequence analysis of PTEN S-glutathionylation sites using RedoxDB

using the DBCP program with a strong probability (72 %) among all the cysteine residues. Crystal structure of PTEN cysteine residues Cys71 and Cys124 oxidized by peroxide has been solved and is available on PDB (ID: 5BUG) demonstrating redox sensitivity of the two residues (Fig. 10).

5. CYSPRED is another neural network-based predictor that determines bonding state of cysteine residues with ~80 % accuracy [42].

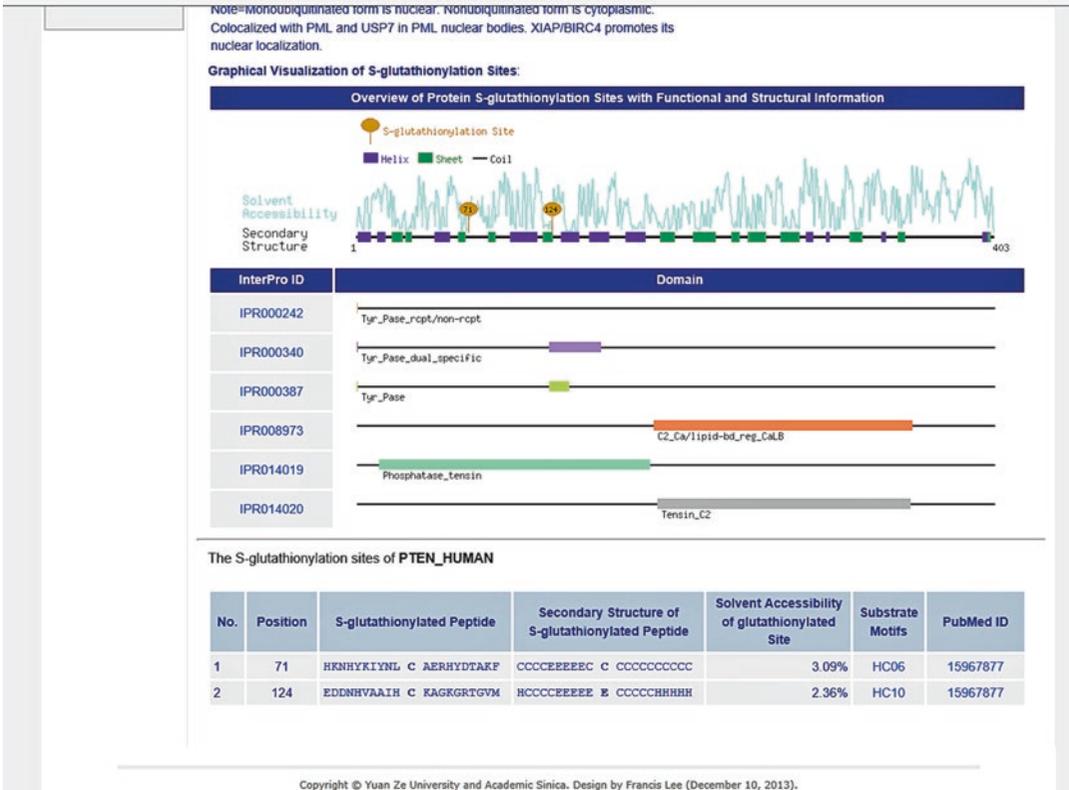


Fig. 9 Analysis of PTEN S-glutathionylation in the dbGSH database

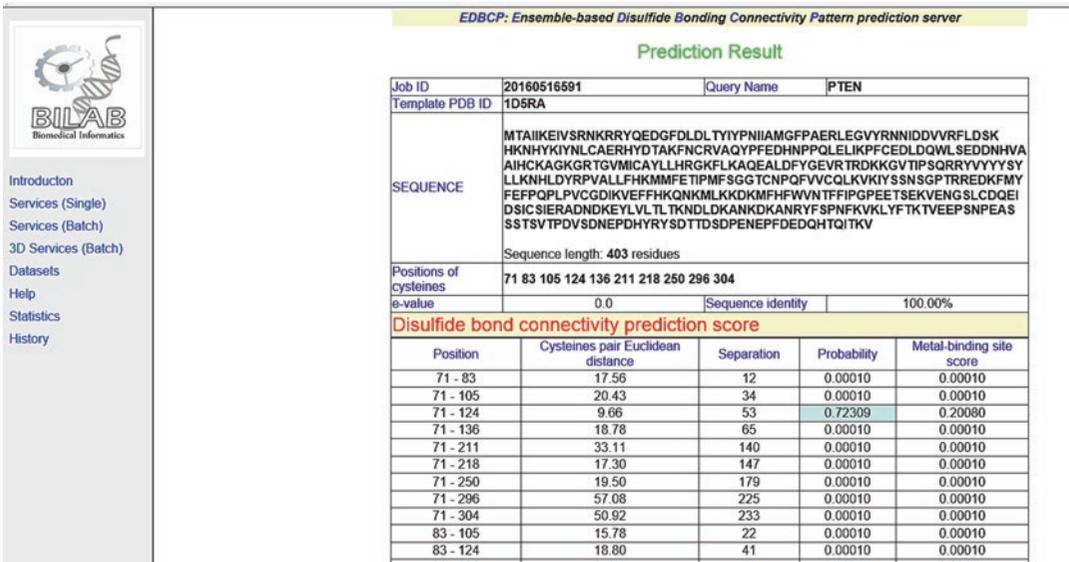


Fig. 10 Disulfide bond formation prediction for PTEN using DBCP program

More experimental data from global redox proteomics studies are required (1) to build databases pertaining to select oxidation states of proteins, (2) to better understand the structural and sequence motifs around redox sensitive versus insensitive cysteine residues, and, (3) to reconcile information from curated databases with known biological properties of protein targets for enabling reliable predictive computational tools for redox active proteins.

3.4 Redox Modifications and Their Impact on Drug Pharmacology

Emerging literature suggests that a significant number of drug targets can be S-modified [2, 5, 6, 28, 29]. Because cysteine residues occur on most proteins [2], redox sensitive residues that impact drug pharmacology need to be identified by harnessing biochemical, cellular, and computational approaches (see previous sections). Unfortunately, the analysis can be complicated by different functional effects of enzyme inhibition and redox biology (i.e., redox modification may affect a subset of a drug target's functions). As such, a solid understanding of the cellular effects driven by the drug target's redox biology and the drug's chemical biology is required before the combination of the two is characterized. Understandably, this field is in its infancy with few examples of impact. One of the most advanced, yet simple example is the receptor tyrosine kinase EGFR because it has a redox sensitive cysteine residue in its active site (Cys₇₉₇) that is also targeted by covalent drugs [5, 19]—this therapeutic strategy is exploited to target other signaling enzymes [43]. Biochemical studies show that redox modifications to the EGFR active site cysteine residue alter the interactions with both reversible and covalent drugs [19]. Recently, an EGFR Cys797Ser mutation has been reported which confers resistance to a covalent drug [44]. This finding highlights S-modification at critical cysteine residues as being able to affect drug pharmacology and was predicted by analyzing the binding interactions in biochemical systems [19]. As redox modifications of cysteine residues can affect conformational dynamics and the binding site topography, analysis of the impact of this PTM on pharmacology can begin by characterizing drug binding to well-characterized redox-modified proteins (*see* Subheading 3.1.1). From these studies, a hypothesis of the biochemical effect of the redox modification on drug interactions can be formulated. Cellular analyses can follow once the targeted protein's redox modification and biophysical effect (e.g., altered drug affinity) are known. The drug effects can be monitored as the protein's redox state is modulated in the cell (*see* Subheading 3.1.2). The redox state of a drug target may impact drug affinity or a subset of its cellular function (i.e., subcellular trafficking). Therefore, the selection of the cellular readout is critical. For example, covalent inhibitors of EGFR (e.g., dacomitinib) block membrane-bound EGFR from initiating a signaling cascade. In contrast, redox modification of EGFR-C₇₉₇ blocks a smaller subset of the EGFR population that

undergoes subcellular trafficking (BWM unpublished result). As such, analyses of membrane-bound EGFR function may miss the redox-regulated EGFR biology. Taken together, the understanding of redox PTM's on drug pharmacology is an interdependent analytical process that requires an advanced understanding of the target's redox biology as well as the drug's chemical biology.

4 Notes

1. Analysis of isolated proteins (*see* Subheading 3.1.1).
For setting up in vitro protein oxidation/S-glutathionylation studies, use 1:250 ratio (protein:GSSG) as a general guideline and always optimize it to specific protein under study. Highly redox sensitive cysteine residues can modify at much lower ratio. An important aspect when setting up in vitro redox modifications is to eliminate presence of any reducing agent (e.g., DTT, TCEP, β -ME) to preserve the modification or label for detection. NEM is relatively more selective for sulfhydryls than IAA and should be preferred for irreversible alkylation or blocking step of cysteine thiol ($-\text{SH}$).
2. *Biotin switch method* (*see* Subheading 3.1.1). A critical aspect for generating high-quality results is to minimize auto- and photo-oxidation of cysteine residues by conducting all treatments under light-limiting conditions (use aluminum foil or incubate in dark at all times) and in degassed buffers. Also, all reagents should be prepared fresh each time to minimize undesirable modification(s).
3. *Chemical biology analysis of cells with redox reagents* (*see* Subheading 3.1.2). Always adjust the culture medium pH after reconstituting the reagent(s) to restore it to a physiological level. Failure to do so can introduce undesirable effects in the cells due to acidic pH.
4. *Redox proteomics analysis* (*see* Subheading 3.2). For SILAC labeling experiments, it is important to remember that dialyzed serum be used instead of complete serum to minimize contamination of unlabeled amino acids with the SILAC culture medium. However, note that some cell lines may not grow well in the presence of dialyzed serum and cell proliferation should be tested with dialyzed serum in regular cell culture medium prior to culturing them in SILAC medium. For obtaining cleaner enrichment, sometimes stringent washes with 25–50 mM NaCl or 1–2 M urea of the affinity captured samples can be considered. However, they should be followed by additional washes with 1 \times lysis buffers to ensure complete removal of the salts. This step can also lead to a certain degree of loss of the desired proteins. Optimizing immunoprecipita-

tion protocols prior to SILAC experiments is strongly suggested to prevent waste of labeled samples. For redox studies, elution of protein from beads should be performed as far as possible by boiling the beads in sample buffer as opposed to using low pH buffer-based elution method to eliminate any potential alterations of the cysteine residues.

References

1. Pace NJ, Weerapana E (2013) Diverse functional roles of reactive cysteines. *ACS Chem Biol* 8(2):283–296. doi:[10.1021/cb3005269](https://doi.org/10.1021/cb3005269)
2. Go YM, Chandler JD, Jones DP (2015) The cysteine proteome. *Free Radic Biol Med* 84:227–245. doi:[10.1016/j.freeradbiomed.2015.03.022](https://doi.org/10.1016/j.freeradbiomed.2015.03.022)
3. Bachi A, Dalle-Donne I, Scaloni A (2013) Redox proteomics: chemical principles, methodological approaches and biological/biomedical promises. *Chem Rev* 113(1):596–698. doi:[10.1021/cr300073p](https://doi.org/10.1021/cr300073p)
4. Poole LB (2015) The basics of thiols and cysteines in redox biology and chemistry. *Free Radic Biol Med* 80:148–157. doi:[10.1016/j.freeradbiomed.2014.11.013](https://doi.org/10.1016/j.freeradbiomed.2014.11.013)
5. Wani R, Nagata A, Murray BW (2014) Protein redox chemistry: post-translational cysteine modifications that regulate signal transduction and drug pharmacology. *Front Pharmacol* 5:224. doi:[10.3389/fphar.2014.00224](https://doi.org/10.3389/fphar.2014.00224)
6. Miki H, Funato Y (2012) Regulation of intracellular signalling through cysteine oxidation by reactive oxygen species. *J Biochem* 151(3):255–261. doi:[10.1093/jb/mvs006](https://doi.org/10.1093/jb/mvs006)
7. Tajc SG, Tolbert BS, Basavappa R, Miller BL (2004) Direct determination of thiol pKa by isothermal titration microcalorimetry. *J Am Chem Soc* 126(34):10508–10509. doi:[10.1021/ja047929u](https://doi.org/10.1021/ja047929u)
8. Gupta V, Carroll KS (2014) Sulfenic acid chemistry, detection and cellular lifetime. *Biochim Biophys Acta* 1840(2):847–875. doi:[10.1016/j.bbagen.2013.05.040](https://doi.org/10.1016/j.bbagen.2013.05.040)
9. Popov D (2014) Protein S-glutathionylation: from current basics to targeted modifications. *Arch Physiol Biochem* 120(4):123–130. doi:[10.3109/13813455.2014.944544](https://doi.org/10.3109/13813455.2014.944544)
10. Jaffrey SR, Snyder SH (2001) The biotin switch method for the detection of S-nitrosylated proteins. *Sci STKE* 2001(86):pl1. doi:[10.1126/stke.2001.86.pl1](https://doi.org/10.1126/stke.2001.86.pl1)
11. Forrester MT, Foster MW, Benhar M, Stamler JS (2009) Detection of protein S-nitrosylation with the biotin-switch technique. *Free Radic Biol Med* 46(2):119–126. doi:[10.1016/j.freeradbiomed.2008.09.034](https://doi.org/10.1016/j.freeradbiomed.2008.09.034)
12. Mustafa AK, Gadalla MM, Sen N, Kim S, Mu W, Gazi SK, Barrow RK, Yang G, Wang R, Snyder SH (2009) H2S signals through protein S-sulfhydration. *Sci Signal* 2(96):ra72. doi:[10.1126/scisignal.2000464](https://doi.org/10.1126/scisignal.2000464)
13. Wani R, Qian J, Yin L, Bechtold E, King SB, Poole LB, Paek E, Tsang AW, Furdulic CM (2011) Isoform-specific regulation of Akt by PDGF-induced reactive oxygen species. *Proc Natl Acad Sci U S A* 108(26):10550–10555. doi:[10.1073/pnas.1011665108](https://doi.org/10.1073/pnas.1011665108)
14. Furdulic CM, Poole LB (2014) Chemical approaches to detect and analyze protein sulfenic acids. *Mass Spectrom Rev* 33(2):126–146. doi:[10.1002/mas.21384](https://doi.org/10.1002/mas.21384)
15. Murray CI, Van Eyk JE (2012) Chasing cysteine oxidative modifications: proteomic tools for characterizing cysteine redox status. *Circ Cardiovasc Genet* 5(5):591. doi:[10.1161/CIRCGENETICS.111.961425](https://doi.org/10.1161/CIRCGENETICS.111.961425)
16. Chang YC, Huang CN, Lin CH, Chang HC, Wu CC (2010) Mapping protein cysteine sulfonic acid modifications with specific enrichment and mass spectrometry: an integrated approach to explore the cysteine oxidation. *Proteomics* 10(16):2961–2971. doi:[10.1002/pmic.200900850](https://doi.org/10.1002/pmic.200900850)
17. Couvreur SM, Zhou Y, Weerapana E (2014) Chemical-proteomic strategies to investigate cysteine posttranslational modifications. *Biochim Biophys Acta* 1844(12):2315–2330. doi:[10.1016/j.bbapap.2014.09.024](https://doi.org/10.1016/j.bbapap.2014.09.024)
18. Chen CA, Wang TY, Varadaraj S, Reyes LA, Hemann C, Talukder MA, Chen YR, Druhan LJ, Zweier JL (2010) S-glutathionylation uncouples eNOS and regulates its cellular and vascular function. *Nature* 468(7327):1115–1118. doi:[10.1038/nature09599](https://doi.org/10.1038/nature09599)
19. Schwartz PA, Kuzmic P, Solowiej J, Bergqvist S, Bolanos B, Almaden C, Nagata A, Ryan K, Feng J, Dalvie D, Kath JC, Xu M, Wani R, Murray BW (2014) Covalent EGFR inhibitor analysis reveals importance of reversible interactions to potency and mechanisms of drug resistance. *Proc Natl Acad Sci U S A* 111:173–178. doi:[10.1073/pnas.1313733111](https://doi.org/10.1073/pnas.1313733111)
20. Le HT, Chaffotte AF, Demey-Thomas E, Vinh J, Friguet B, Mary J (2009) Impact of hydrogen

- peroxide on the activity, structure, and conformational stability of the oxidized protein repair enzyme methionine sulfoxide reductase A. *J Mol Biol* 393(1):58–66. doi:[10.1016/j.jmb.2009.07.072](https://doi.org/10.1016/j.jmb.2009.07.072)
21. Zmijewski JW, Banerjee S, Bae H, Friggeri A, Lazarowski ER, Abraham E (2010) Exposure to hydrogen peroxide induces oxidation and activation of AMP-activated protein kinase. *J Biol Chem* 285(43):33154–33164. doi:[10.1074/jbc.M110.143685](https://doi.org/10.1074/jbc.M110.143685)
 22. Victor KG, Rady JM, Cross JV, Templeton DJ (2012) Proteomic profile of reversible protein oxidation using PROP, purification of reversibly oxidized proteins. *PLoS One* 7(2):e32527. doi:[10.1371/journal.pone.0032527](https://doi.org/10.1371/journal.pone.0032527)
 23. Kim JR, Yoon HW, Kwon KS, Lee SR, Rhee SG (2000) Identification of proteins containing cysteine residues that are sensitive to oxidation by hydrogen peroxide at neutral pH. *Anal Biochem* 283(2):214–221. doi:[10.1006/abio.2000.4623](https://doi.org/10.1006/abio.2000.4623)
 24. Sethuraman M, McComb ME, Huang H, Huang S, Heibeck T, Costello CE, Cohen RA (2004) Isotope-coded affinity tag (ICAT) approach to redox proteomics: identification and quantitation of oxidant-sensitive cysteine thiols in complex protein mixtures. *J Proteome Res* 3(6):1228–1233. doi:[10.1021/pr049887e](https://doi.org/10.1021/pr049887e)
 25. Chiappetta G, Ndiaye S, Igbaria A, Kumar C, Vinh J, Toledano MB (2010) Proteome screens for Cys residues oxidation: the redoxome. *Methods Enzymol* 473:199–216. doi:[10.1016/S0076-6879\(10\)73010-X](https://doi.org/10.1016/S0076-6879(10)73010-X)
 26. Weerapana E, Wang C, Simon GM, Richter F, Khare S, Dillon MB, Bachovchin DA, Mowen K, Baker D, Cravatt BF (2010) Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature* 468(7325):790–795. doi:[10.1038/nature09472](https://doi.org/10.1038/nature09472)
 27. Kim HJ, Ha S, Lee HY, Lee KJ (2015) ROSics: chemistry and proteomics of cysteine modifications in redox biology. *Mass Spectrom Rev* 34(2):184–208. doi:[10.1002/mas.21430](https://doi.org/10.1002/mas.21430)
 28. Paulech J, Solis N, Edwards AV, Puckeridge M, White MY, Cordwell SJ (2013) Large-scale capture of peptides containing reversibly oxidized cysteines by thiol-disulfide exchange applied to the myocardial redox proteome. *Anal Chem* 85(7):3774–3780. doi:[10.1021/ac400166e](https://doi.org/10.1021/ac400166e)
 29. Yuan K, Liu Y, Chen HN, Zhang L, Lan J, Gao W, Dou Q, Nice EC, Huang C (2015) Thiol-based redox proteomics in cancer research. *Proteomics* 15(2–3):287–299. doi:[10.1002/pmic.201400164](https://doi.org/10.1002/pmic.201400164)
 30. Mann M (2006) Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol* 7(12):952–958. doi:[10.1038/nrm2067](https://doi.org/10.1038/nrm2067)
 31. Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* 1(6):2856–2860. doi:[10.1038/nprot.2006.468](https://doi.org/10.1038/nprot.2006.468)
 32. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlett-Jones M, He F, Jacobson A, Pappin DJ (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 3(12):1154–1169. doi:[10.1074/mcp.M400129-MCP200](https://doi.org/10.1074/mcp.M400129-MCP200)
 33. Leichert LI, Gehrke F, Gudiseva HV, Blackwell T, Ilbert M, Walker AK, Strahler JR, Andrews PC, Jakob U (2008) Quantifying changes in the thiol redox proteome upon oxidative stress in vivo. *Proc Natl Acad Sci U S A* 105(24):8197–8202. doi:[10.1073/pnas.0707723105](https://doi.org/10.1073/pnas.0707723105)
 34. Marino SM, Gladyshev VN (2011) Redox biology: computational approaches to the investigation of functional cysteine residues. *Antioxid Redox Signal* 15(1):135–146. doi:[10.1089/ars.2010.3561](https://doi.org/10.1089/ars.2010.3561)
 35. Raimondi D, Orlando G, Vranken WF (2015) Clustering-based model of cysteine co-evolution improves disulfide bond connectivity prediction and reduces homologous sequence requirements. *Bioinformatics* 31(8):1219–1225. doi:[10.1093/bioinformatics/btu794](https://doi.org/10.1093/bioinformatics/btu794)
 36. Sun MA, Wang Y, Cheng H, Zhang Q, Ge W, Guo D (2012) RedoxDB—a curated database for experimentally verified protein oxidative modification. *Bioinformatics* 28(19):2551–2552. doi:[10.1093/bioinformatics/bts468](https://doi.org/10.1093/bioinformatics/bts468)
 37. Bostan H, Salim N, Hussein ZA, Klappa P, Shamsir MS (2012) CMD: a database to store the bonding states of cysteine motifs with secondary structures. *Adv Bioinf* 2012:849830. doi:[10.1155/2012/849830](https://doi.org/10.1155/2012/849830)
 38. Chen YJ, Lu CT, Lee TY, Chen YJ (2014) dbGSH: a database of S-glutathionylation. *Bioinformatics* 30(16):2386–2388. doi:[10.1093/bioinformatics/btu301](https://doi.org/10.1093/bioinformatics/btu301)
 39. Ferre F, Clote P (2005) DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res* 33(Web Server issue):W230–W232. doi:[10.1093/nar/gki412](https://doi.org/10.1093/nar/gki412)
 40. Ceroni A, Passerini A, Vullo A, Frasconi P (2006) DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res* 34(Web Server issue):W177–W181. doi:[10.1093/nar/gkl266](https://doi.org/10.1093/nar/gkl266)
 41. Lin HH, Tseng LY (2010) DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. *Nucleic Acids Res*

- 38(Web Server issue):W503–W507. doi:[10.1093/nar/gkq514](https://doi.org/10.1093/nar/gkq514)
42. Fariselli P, Riccobelli P, Casadio R (1999) Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins* 36(3):340–346
43. Visscher M, Arkin MR, Dansen TB (2015) Covalent targeting of acquired cysteines in cancer. *Curr Opin Chem Biol* 30:61–67. doi:[10.1016/j.cbpa.2015.11.004](https://doi.org/10.1016/j.cbpa.2015.11.004)
44. Yu HA, Tian SK, Drilon AE, Borsu L, Riely GJ, Arcila ME, Ladanyi M (2015) Acquired resistance of EGFR-mutant lung cancer to a T790M-specific EGFR inhibitor: emergence of a third mutation (C797S) in the EGFR tyrosine kinase domain. *JAMA Oncol* 1(7):982–984. doi:[10.1001/jamaoncol.2015.1066](https://doi.org/10.1001/jamaoncol.2015.1066)

Chapter 10

Analysis of Protein Phosphorylation and Its Functional Impact on Protein–Protein Interactions via Text Mining of the Scientific Literature

Qinghua Wang, Karen E. Ross, Hongzhan Huang, Jia Ren, Gang Li, K. Vijay-Shanker, Cathy H. Wu, and Cecilia N. Arighi

Abstract

Post-translational modifications (PTMs) are one of the main contributors to the diversity of proteoforms in the proteomic landscape. In particular, protein phosphorylation represents an essential regulatory mechanism that plays a role in many biological processes. Protein kinases, the enzymes catalyzing this reaction, are key participants in metabolic and signaling pathways. Their activation or inactivation dictate downstream events: what substrates are modified and their subsequent impact (e.g., activation state, localization, protein–protein interactions (PPIs)). The biomedical literature continues to be the main source of evidence for experimental information about protein phosphorylation. Automatic methods to bring together phosphorylation events and phosphorylation-dependent PPIs can help to summarize the current knowledge and to expose hidden connections. In this chapter, we demonstrate two text mining tools, RLIMS-P and eFIP, for the retrieval and extraction of kinase–substrate–site data and phosphorylation-dependent PPIs from the literature. These tools offer several advantages over a literature search in PubMed as their results are specific for phosphorylation. RLIMS-P and eFIP results can be sorted, organized, and viewed in multiple ways to answer relevant biological questions, and the protein mentions are linked to UniProt identifiers.

Key words Bioinformatics, Phosphorylation, Post-translational modification, Protein–protein interaction, Text mining

1 Introduction

Post-translational modifications (PTMs) are an important contributor to protein diversity. PTMs play a pivotal role in protein function, regulating activity, localization, and protein–protein interactions (PPIs), and therefore disruptions in PTMs can lead to disease [1]. In particular, protein phosphorylation is an essential regulatory mechanism in many biological processes. Proteins can be phosphorylated at different and/or multiple positions, most commonly on serine, threonine, and tyrosine residues. Protein kinases, the enzymes catalyzing the phosphorylation reaction, play

a key role in regulating these events and have become therapeutic targets for drug design in multiple diseases [2–4]. However, few drugs targeting kinases have been completely successful in the clinic mainly due to the conserved nature of kinases. Consequently, many of the available inhibitors lack sufficient selectivity for effective clinical application. The identification and characterization of kinase–substrate interactions are keys to improve the approaches to targeted drug development [5].

The scientific literature contains a wealth of protein phosphorylation data derived both from traditional experiments that focus on a small number of proteins and from high-throughput experiments that attempt to assess the phosphorylation state of the whole proteome [6]. Researchers frequently query PubMed or specialized databases to gain access to this information. Similarly, database biocurators collect literature, and read and extract the most salient information relevant to their domain. Given the continuing increase of the size of the PubMed database, finding or collecting information that is spread across this vast knowledge pool remains challenging. Automatic methods to bring this data together can help to summarize the current knowledge and to expose hidden connections. For example, one article might describe that phosphorylation of a protein at a given site is implicated in a particular disease, and another article might describe a kinase that phosphorylates the site, leading to the connection of the kinase to the disease, which could be investigated further. Text mining tools have evolved considerably in number and quality and are being used to address a variety of research questions in the biomedical domain; for recent reviews see [7–9].

*i*ProLINK (integrated Protein Literature, Information and Knowledge) [10] offers a portfolio of text mining tools and annotated corpora developed by our group. Some of these are intended for developers to serve as modules in specific steps of their text mining pipelines (e.g., *i*XtractR [11] for relation extraction, and *i*Simp for sentence simplification [12]). Others are applications for biomedical researchers and biocurators to facilitate the exploration of the literature about proteins (pGenN [13], eFIP [14, 15], eGIFT [16], and RLIMS-P [17, 18]) and microRNAs (miRTex [19]) (Table 1).

Among these applications, RLIMS-P and eFIP facilitate the extraction of phosphorylation information from the literature and therefore are the focus of this book chapter.

RLIMS-P is a rule-based information extraction system that identifies kinase, substrate, and site relations in the scientific literature (including PubMed abstracts and PMC open access (OA) full-length articles). For example, the tuple <Akt, CHK1, Ser280> is extracted by RLIMS-P from the following sentence:

“CHK1 is directly phosphorylated by Akt at Ser280, a modification that results in cytoplasmic sequestration” [20].

Table 1
Text mining tools available in iProLINK

Tool	Description	Bioentities/relations	Standard used
pGenN	Identifies plant gene name mentions in Medline abstracts	<ul style="list-style-type: none"> • Protein/gene 	<ul style="list-style-type: none"> • UniProt identifier • EntrezGene
eGIFT	Identifies informative terms (<i>i</i> Terms) and documents relevant to a gene/protein (abstract level)	<ul style="list-style-type: none"> • Protein/gene • Informative term (<i>i</i>Term) 	<ul style="list-style-type: none"> • GO term • UniProt Keyword
miRTex	Identifies miRNA-target relations as well as miRNA-gene and gene-miRNA regulation relations in Medline abstracts	<ul style="list-style-type: none"> • miRNA-target • gene-miRNA • miRNA-gene 	
RLIMS-P	Identifies information relevant to protein phosphorylation: kinase, substrate, and sites. (abstract and full-length PMC open access articles)	<ul style="list-style-type: none"> • Kinase-substrate • Substrate-site • Kinase-substrate-site 	<ul style="list-style-type: none"> • UniProt Identifier
eFIP	Identifies phosphorylation-dependent protein-protein interactions (abstract and full-length PMC open access articles)	<ul style="list-style-type: none"> • Phosphorylation-dependent PPI • Impact on PPI (promote or inhibit) 	<ul style="list-style-type: none"> • UniProt Identifier

Since these three entities (kinase, substrate, and site) are rarely comentioned in the same sentence, RLIMS-P employs techniques that combine information found in different sentences. The kinase or substrate names detected could correspond to individual proteins (e.g., Crm1), protein complexes (e.g., CDK1-cyclin-B), or a group of related proteins (e.g., Src kinases), whereas a site could be a residue type (e.g., serine, threonine, and tyrosine), a specific residue (e.g., Ser-391), or a protein region or domain (e.g., C-terminal domain) [18]. RLIMS-P has been benchmarked with multiple corpora [17]. The F-scores (harmonic mean between precision and recall), based on a collection of sections derived from 100 full-text articles, have previously been reported to be 0.88, 0.91, and 0.92 for kinases, substrates, and sites, respectively [17]. In addition, RLIMS-P integrates GNormPlus [21] to link the detected kinase and substrate names to UniProt identifiers whenever possible.

eFIP builds on RLIMS-P by first detecting mentions of protein phosphorylation (kinase, substrate, and site), but adds detection of protein-protein interactions (PPIs) involving the phosphorylated protein. The types of PPIs captured include interactions between two proteins, or interactions between a protein and a protein complex, protein region, or protein class. Once the phosphorylation and PPI mentions are detected, the second step is to identify a possible relation between the two events. The evaluation of eFIP on full-length articles achieved an F-measure of 0.84 on 100 article

sections [14]. Selected data from RLIMS-P and eFIP has been integrated in iPTMnet (<http://proteininformationresource.org/iPTMnet/>) and is actively used in the curation of proteoforms in the Protein Ontology [22].

This chapter demonstrates how to use RLIMS-P and eFIP to uncover information about protein phosphorylation and phosphorylation-dependent PPIs from the literature.

2 Materials

2.1 Web Sites

iProLINK: <http://proteininformationresource.org/iprolink>

RLIMS-P: <http://proteininformationresource.org/rlimsp>

eFIP: <http://proteininformationresource.org/efip>

2.2 General Aspects of the RLIMS-P and eFIP Interfaces

Input: Both the RLIMS-P and eFIP web sites allow the input of keywords or phrases that can be combined with Boolean operators (AND, OR, NOT) in the same way as building a PubMed query. Similarly, MeSH terms (controlled vocabulary used to index Medline abstracts) can be included in the search (e.g., “Alzheimer Disease”[Mesh]). The input is sent to the PubMed web site and relevant PMIDs are retrieved. The PMIDs are then used to query a backend database that hosts preprocessed results for PubMed abstracts and full-length PMC OA documents by RLIMS-P or eFIP. In both systems, you have the option to restrict the search to a particular organism of interest (Fig. 1a 3, Fig. 2a). You can also select to exclude review articles if you are only interested in research articles, and/or query only abstracts (Fig. 1a 4). eFIP also supports searches based on protein roles (kinases, substrates, interacting partners) for protein names. Alternatively, a list of PMIDs or PMCID, delimited by comma, space, or listed in new lines, can be entered (Fig. 1a 5, Fig. 2a).

Results: The RLIMS-P result page presents summary statistics of the retrieved results (Fig. 1b 1), listing separately the number of documents with potential phosphorylation information (i.e., those with the word “phosphorylation” or similar ones) and those with phosphorylation information according to RLIMS-P (i.e., there is at least one substrate identified). In addition, eFIP shows the summary statistics for interactants detected (Fig. 2b 1).

Editing capabilities: To unlock editing capabilities, user registration and login are required (Fig. 1a 1, Fig. 2a, *see Note 1*). Edited results can be downloaded.

Cytoscape: eFIP offers a graphical view of the text mining results, displaying the protein entities as nodes and their relations as edges. The node names correspond to the protein entities in the result table, with some of the longer names abbreviated. The graph can be

A

1 Login

RLIMS-P: Rule-based Literature Mining System for Protein Phosphorylation



RLIMS-P is a rule-based text-mining program specifically designed to extract protein phosphorylation information on protein kinase, substrate and phosphorylation sites from biomedical literature (Torii *et al.*, 2015). RLIMS-P currently works on PubMed abstracts and open access full text articles.

- Web Service
- RLIMS-P 2.0 Corpus
- Supplementary Material

RLIMS-P Search Form

Enter Keywords (accepts Boolean operators (AND, OR, NOT)) 2 3
 CHK1 OR CHEK1 OR "checkpoint kinase-1" 4 Restrict by organism Submit Query Reset

Exclude review papers Only abstracts

Or Enter PubMed IDs (PMIDs) delimited by "," or space, e.g., 15234272, 16436437. 5
 Input PMID

Submit Query Reset

You can process up to 200 PMIDs per run. [Sample output](#)

B

1266 documents with potential phosphorylation are processed [Save PMIDs](#)
 Documents with phosphorylation mentions=1162 where PTM enzyme (kinase)=278, Substrate=854, Site=245 1
 Click [here](#) to see results of the latest 200 PMID. ?

Summary

Show all annotations

2 View by Summary

Download

Help

Show Selected	PubMed ID	PTM enzyme	Phosphorylated Protein (Substrate)	No. of Sentences	Text Evidence
<input type="checkbox"/>	22977173 PMC3510507	dna-pk, atr, rpa32, dna-pkcs, dna-pk (29), dna-pk (figure 2a-c), atm, pikks/cdk, pikl	rpa32, wt, rad17, kap1, chk2 checkpoint proteins, chk1, s4a/s8a mutant, ps10-h3, dna-pkcs, ser345, ser317, chk1 ser345, toppb1s atr, chk2, h2ax, rpa, mre11, toppb1	23	🔗

3

View by Substrate

Show all annotations

View by Substrate

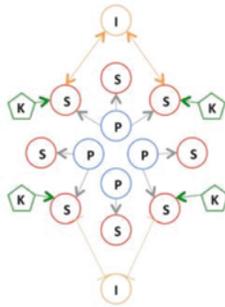
5 Download

Help

Show All	Phosphorylated Protein (Substrate)	PubMed ID	PTM enzyme	Phosphorylation Site	No. of Sentences	Text Evidence
<input type="checkbox"/>	chk1	23748345	pim	Ser-280	1	🔗 4
<input type="checkbox"/>		24663817 PMC3963969	rad3	Thr-11, Thr-645, Ser-3, Thr-412	1	🔗
<input type="checkbox"/>		20798862 PMC2925474	atm/atr	Ser, Ser-345	1	🔗
<input type="checkbox"/>		20798862 PMC2925474	cdk1	Ser-286, Ser-301	1	🔗
<input checked="" type="checkbox"/>		23748345	akt	Ser-280	1	🔗
<input type="checkbox"/>		23748345	pim kinases		1	🔗

Fig. 1 RLIMS-P for extraction of kinase–substrate–site information about CHK1. (a) RLIMS-P home page, showing the different functionalities: login capability (1), input query options such as keywords (2) or PubMed IDs (5) and search options (3, organism restriction or 4, exclusion of review articles or only abstracts). (b) Partial display of RLIMS-P results for CHK1 search with summary statistics (1), and tables with “View by Summary” (2), and “View by Substrate” views (3). The “Text Evidence” (4) column provides links to the text evidence page. Text mining results can be downloaded in CSV format (5)

A



eFIPonline is a website to assist in the retrieval of relevant articles describing protein phosphorylation-dependent protein-protein interactions (PPIs), together with extraction of information about protein kinases, substrates, and interacting partners. Both Medline abstracts and PMC full-length articles are displayed. Results are pre-processed and updated on a quarterly basis. These can be queried by searching for proteins in any of the three types (kinase, substrate, or interacting partner), searching directly for PMIDs/PMCID, or by using a PubMed-style query.

Enter Keywords (accepts Boolean operators AND, OR, NOT) ¹

e.g., apoptosis AND Bad AND phosphorylation

Submit

Reset

Restrict by organism

Enter Protein Names and Type (accepts Boolean operator OR for synonyms) ²

CHK1 OR CHECK OR "checkpoint kinase-1"

All Types

Submit

All Types

Kinase

Substrate

Interactant

Enter a List of IDs (delimited by comma or space) and **Specify the Type**

e.g., 15234272, 16436437

PMID

Submit

Reset

B

Documents with potential phosphorylation=**867** [Save PMIDs](#)

Documents with phosphorylation mentions=**856** where Kinase=**462**, Substrate=**1285**, Site=**398**, Interacting Protein=**127** and phosphorylation → PPI=**231** ¹

Summary

(Download Table) (Cytoscape View) ⁵

Show all annotations ⁺

View by Summary ²

Show Selected	PubMed ID	Protein Kinase	Phosphorylated Protein (Substrate)	Interactant	No. of Sentences	Text Evidence
<input type="checkbox"/>	17480229 PMC1868713	cdc2/cyclin b, chk1, cdk2, erk-map kinase, myt1, wee1, ck2, cdc2, cyclin -dependent kinase	cdc25, ser287, pp1s, xenopus wee1, cdc25s, 14-3-3, cdc2, wee1, b56delta, thr138, cdc2/cyclin b -mediated phosphorylation, ser123, cdc2/cyclin b, specific antibody, cdc25c, okadaic acid, the human and xenopus wee1 kinases	pp1, 14-3-3, pp1s, cdc25, mitotic cdc25, 14-3-3s, cyclin b, pin1, myt1	3 14	4 🔗

Fig. 2 eFIP for extraction of phosphorylation-dependent PPI information about CHK1. (a) eFIP home page, showing the different functionalities: input query options such as keywords (1), protein names (2), or PubMed/PMC IDs (4) and search options for organism restriction and protein type including substrate, kinase or interactant (3). (b) Partial display of eFIP results for CHK1 search with summary statistics (1), and table with “View by Summary” (2). The columns “No. of Sentences” (3) and “Text Evidence” (4) provide links to the text evidence pages. Text mining results can be downloaded in CSV format and be viewed in Cytoscape (5)

saved in PNG and XGMML-beta (Cytoscape compatible) format (Fig. 6 2). Substrates, kinases, and interactants are represented as nodes with red circles, green pentagons, and orange circles, respectively. Interactions that are enabled or enhanced by phosphorylation are depicted as edges using solid orange lines with pointed arrowheads, whereas those that are decreased or inhibited are depicted by dashed orange lines with T-type arrowheads (Fig. 6).

3 Methods

For illustration purposes, we will showcase RLIMS-P and eFIP tool usage with examples from the Checkpoint kinase-1 protein, commonly referred to as CHK1 or CHEK1. This protein is a serine-/threonine-specific protein kinase. It coordinates the DNA damage response (DDR) and cell cycle checkpoint response [23]. Activation of CHK1 results in the initiation of cell cycle checkpoints, cell cycle arrest, DNA repair, and cell death to prevent damaged cells from progressing through the cell cycle [24]. A recent review article by Goto *et al.* [25] describes the regulation of CHK1 via phosphorylation, its substrates, and the functional impact. To validate the approach, we compare the output of our text mining tools with the knowledge in the review article when applicable. We illustrate in the following text a variety of examples of RLIMS-P and eFIP usage via specific biological questions.

3.1 How to Find Kinases Acting on a Given Substrate. What Sites Are Phosphorylated?

Is CHK1 phosphorylated? If so, which sites? By what kinases? To answer these questions, we will use the RLIMS-P web site (<http://proteininformationresource.org/rlimsp>, Fig. 1a). The goal in this case is to find the articles mentioning CHK1 as a substrate, as we are interested in its phosphorylation sites. To achieve the most comprehensive result, it is recommended to include the different names by which CHK1 is known (e.g., CHEK1, Checkpoint kinase-1). If you are not familiar with the variety of names that are used for your protein of interest, you can check in a reference curated source, such as UniProt [26] or Entrez [27]. For this case, we will use the query (Fig. 1a 2):

CHK1 OR CHEK1 OR “checkpoint kinase-1”

1. *Go to RLIMS-P web site and enter this query in the box and submit.* Results are returned as shown in Fig. 1b. Information on the top of the page summarizes the general statistics for the search results (Fig. 1b I), including the number of articles with potential protein phosphorylation mentions and the number of kinase, substrate, and site mentions (*see Note 2*).
2. *Display results by “Substrate.”* The results from the search in RLIMS-P include articles where the keywords are mentioned and which are about protein phosphorylation. The default table

view is a summary listing the kinase and substrate mentions for each PMID. To obtain the subset where CHK1 is the phosphorylated protein, choose the option “View by Substrate” from the pull-down menu (Fig. 1b 2) (*see Note 3*).

3. *Find CHK1 as substrate.* The table in Fig. 1b 3 is now substrate centric. Next, we have to find CHK1 in the substrate column. As shown in this table, there are many articles describing phosphorylation of CHK1 (where CHK1 acts as a substrate). In addition, the kinases that phosphorylate CHK1 and the phosphorylation sites can now be easily identified in the columns “PTM enzyme” and “phosphorylation site,” respectively.
4. *Validate and summarize the information.* When the results are viewed by substrate (as shown in Fig. 1b 3), all the phosphorylation sites on a substrate are shown. Now continue with our example by looking for CHK1 as substrate. The “No. of Sentences” column provides quick access to evidence sentences with color-coded highlighting of kinase (*green*), substrate (*blue*), and site (*red*) mentions (*see Fig. 4* bottom panel). This page is almost the same as the page linked out through icons in the “Text Evidence” column (Fig. 1b 4), except that it restricts its sentence display to those where the information tuples are directly derived. To validate the information, the evidence can also be viewed by clicking on the icon in the “Text Evidence” column (Fig. 1b 4), which will take you to the evidence page (Fig. 3a). The evidence page presents a table summarizing the data extracted from the article with links to the source sentences (Fig. 3a 2), a block showing the relevant sentences from the text (abstract or full text) with color-coding highlighting (Fig. 3a 3), and the normalization table, which suggests UniProt identifiers for the kinases and substrates detected (Fig. 3a 3–4). Results can be filtered by specific sections of the article (e.g., figure legends, result section, abstract, etc., *see Fig. 3a 1*). If a user is logged in, he or she can validate individual information tuples by clicking on the check or “X” next to the annotation to agree or disagree, respectively (Fig. 3b 1). The example shown in Fig. 3b demonstrates the agreement on data extracted for phosphorylation of Ser-280 on Chk1 by PIM kinases. User can add additional information in the comment box, in this case, the more specific kinase PIM1 (Fig. 3b). In addition, the “Add Annotation” (Fig. 3b 2) allows addition of manually curated information tuples. Furthermore, the normalization table becomes editable after user logs in (Fig. 3b 3–4).

Another way to review the RLIMS-P results is to download them in CSV format, which could be done on a single article or on the selected collective result by clicking the Download button in the right corner of the Results page (Fig. 1b 5). The file can be opened in Excel (Fig. 3c) where you can filter or sort the informa-

A

Previous Page RLIMS-P Home

Text Evidence Choose a specific section: All 1

23748345 2014 L L Yuan, A S Green, S Bertoli, F Grimal, Leukemia Full Text

No.	PTM enzyme	Substrate	Site	Sentence
1	akt	chk1 (f3-td -dependent chk1)	Ser-280	2 (Abstract 1)
2	pim	chk1	Ser-280	10 (Abstract 1)
3	pim kinases	chk1		1 (Abstract 1)
4		histone h3		9 (Abstract 1)
5		cdc25c		8 (Abstract 1)

Gene Normalization 4

Protein	Name	UniProtKB AC	Annotation No.
PTM enzyme	akt	P31749AKT1_HUMAN B0LPE5BOLPE5_HUMAN B3KVH483VH4_HUMAN	1
	pim	P11308PIM1_HUMAN	2
	chk1	O14757CHK1_HUMAN B4DT73B4DT73_HUMAN	2, 3
Substrate	f3-td -dependent chk1	O14757CHK1_HUMAN B4DT73B4DT73_HUMAN	1
	cdc25c	P93007MPP3_HUMAN	5

Abstract (ABSTRACT 1) 3

1 Pim kinases phosphorylate Chk1 and regulate its functions in acute myeloid leukemia .

2 Phosphorylation by Akt on Ser 280 was reported to induce cytoplasmic retention and inactivation of CHK1 with consequent genetic instability in PTEN-/- cells .

3 In acute myeloid leukemia cells carrying the FLT3-internal tandem duplication (ITD) mutation , we observed high rates of FLT3-ITD -dependent CHK1 Ser 280 phosphorylation .

4 Pharmacological inhibition and RNA interference identified Pim1/2 , not Akt , as effectors of this phosphorylation .

5 Pim1 catalyzed Ser 280 phosphorylation in vitro and ectopic expression of Pim1/2 -induced CHK1 phosphorylation .

6 Ser 280 phosphorylation did not modify CHK1 localization , but facilitated its cell cycle and resistance functions in leukemic cells .

...

8 Consistently , etoposide-induced CHK1 -dependent phosphorylations of CDC25C on Ser 216 and histone H3 on Thr11 were decreased upon FLT3 inhibition .

9 Accordingly , ectopic expression of CHK1 improved the resistance of FLT3-ITD cells and maintained histone H3 phosphorylation in response to DNA damage , whereas expression of unphosphorylated Ser 280Ala mutant did not .

10 Finally , FLT3 - and Pim -dependent phosphorylation of CHK1 on Ser 280 was confirmed in primary blasts from patients .

Select/deselect kinase substrate site phospho.keywords

B

No.	PTM enzyme	Substrate	Site	Sentence	Comment	Validation
1	akt	chk1 (f3-td -dependent chk1)	Ser-280	2 (Abstract 1)		<input type="checkbox"/> <input type="checkbox"/>
2	pim	chk1	Ser-280	10 (Abstract 1)	PIM1	<input checked="" type="checkbox"/> <input type="checkbox"/>
3	pim kinases	chk1		1 (Abstract 1)		<input checked="" type="checkbox"/> <input type="checkbox"/>
4		histone h3		9 (Abstract 1)		<input checked="" type="checkbox"/> <input type="checkbox"/>
5		cdc25c		8 (Abstract 1)		<input checked="" type="checkbox"/> <input type="checkbox"/>

Add Annotation 2

Protein	Name	Type	UniProtKB AC	Add UniProtKB AC	Annotation No.
PTM enzyme	akt	Q LipPhos	P31749AKT1_HUMAN	<input type="text"/>	1
		Family	B0LPE5BOLPE5_HUMAN	<input type="text"/>	
		Complex	B3KVH483VH4_HUMAN	<input type="text"/>	
pim		Protein	P11308PIM1_HUMAN	<input type="text"/>	2
		Family		<input type="text"/>	
		Complex		<input type="text"/>	
pim kinases		Protein	Not normalized	<input type="text"/>	3
		Family		<input type="text"/>	
		Complex		<input type="text"/>	
chk1		Protein	O14757CHK1_HUMAN	<input type="text"/>	2, 3
		Family	B4DT73B4DT73_HUMAN	<input type="text"/>	
		Complex		<input type="text"/>	
f3-td -dependent chk1		Protein	O14757CHK1_HUMAN	<input type="text"/>	1
		Family	B4DT73B4DT73_HUMAN	<input type="text"/>	
		Complex		<input type="text"/>	
cdc25c		Protein	P93007MPP3_HUMAN	<input type="text"/>	5
		Family		<input type="text"/>	
		Complex		<input type="text"/>	
histone h3		Protein	Not normalized	<input type="text"/>	4
		Family		<input type="text"/>	
		Complex		<input type="text"/>	

Add Gene Normalization

C

Substrate	PMID	PTM enzyme	Site	Sentence
chk1	20798862	atm/atr	Ser,Ser-345	AB - During DDR such as that resulting from exposure t
chk1	22686412	cdk1	Ser-286,Ser-301	A reduced electrophoretic mobility of Chk1in mitotic c
chk1	20798862	cdk1	Ser-286,Ser-301	As recently reported , during mitosis Cdk1 phosphoryl:
chk1	25024738	atr	Ser-317	H - ATR phosphorylates Chk1 S317 upon replication str
chk1	17210576	vpr	Ser-345	Conversely , H2AX down-regulation had little effect on
chk1	20976184	atm/atr	Ser-345	In addition to phosphorylating H2AX , ATM/ATR also p
chk1	20639859	atr	Ser-296	Ser296 phosphorylation is catalysed by Chk1 itself afte
chk1	20609246	vpr	Ser-345	The S phase -dependent requirement of Chk1-Ser345 ;

Fig. 3 Analysis of CHK1 phosphorylation text evidence for PMID:23748345. (a) RLIMS-P text evidence view. The information can be filtered by the different sections of the article when applicable (1). The table shows kinase-substrate-site data, with each row displaying a unique information tuple with the sentence number and section source (2). The text panel on the right (3) contains the evidence text with the sentence numbers. The kinase, substrate, and site are color coded. The gene normalization table (4) shows possible UniProt identifiers for the kinases and substrates mentioned in the table. (b) RLIMS-P table view when editing capability is unlocked. New columns appear: "Comment" for adding notes, and "Validation" for accepting/rejecting the annotation (1). Missing annotations can be added (2). Normalization data can be validated as well (3). If needed an auto-filled UniProt search can be triggered by clicking on the search icon with "UniProt" link in the "Name" column (this icon appears after hovering over anywhere in a given row) (4). Clicking on the link leads to the corresponding UniProt entries [26]. (c) Partial view of the downloaded CSV format file. The file includes PMID, substrate, kinase, site, and evidence sentence

Table 2
CHK1 phosphorylation sites with kinases validated from RLIMS-P results (species nonspecific)

Site	Kinase	PMIDs
Ser-280	P90 RSK AKT PIM1	19406993, 15710331, 22481935, 15107605, 12062056, 22357623, 23748345
Ser-286	CDK1 CDK2	20798862, 19837665, 22686412, 18983824, 16629900
Ser-296	CHK1	20639859, 22357623, 22686412, 23068608, 20053762
Ser-301	CDK1 CDK2	20798862, 19837665, 22686412, 18983824, 16629900
Ser-317	ATR ATM	21730979, 19625493, 20798862, 18723495, 20062519, 16629900, 16547171
Ser-345	ATR ATM	21289283, 20798862, 20976184, 15107605, 15159397, 22357623, 16629900, 22357623, 23383325, 16547171, 23422000, 11687578, 20053762, 17210576, 20609246

tion as needed. For example, you can download all results and then filter to (1) show those where CHK1 is the substrate and (2) hide rows with “Blank” information in the “PTM enzyme” and “Site” columns (Fig. 3c). The file contains the evidence sentences to assist you in validating the results (*see Note 3*).

The results can be summarized as in Table 2. RLIMS-P found all the sites and kinases cited in the review by Goto *et al.* [25], and in addition, RLIMS-P found an article describing a kinase not listed in that review, namely PIM1 (bold in Table 2).

3.2 How to Find All Substrates for a Given Kinase

Because CHK1 is itself a kinase, we can easily identify all substrates of CHK1 by choosing the “View by Kinase” option (Fig. 4). A variety of substrates are identified here under the column “Phosphorylated Protein (Substrate).” This column can be sorted using the arrow next to the title “Phosphorylation Protein (Substrate)” so that the information regarding the same substrate is brought together. Table 3 shows the summary of substrates of CHK1 and detected phosphorylation sites. Based on the number of articles linked to the substrates, CDC25 proteins seem to be the most widely studied CHK1 substrates.

3.3 How to Find the Interacting Partners of Phosphorylated Proteins

In our examples in the following text, we address (1) how phosphorylation on CHK1 affects its interaction with other proteins, (2) how PPIs are affected by proteins phosphorylated by CHK1, and (3) how phosphorylation of other proteins affect their interaction with CHK1. eFIP is capable of identifying the impact of phosphorylation, e.g., whether the phosphorylation enables the binding to a partner or inhibits the binding.

View by Kinase Show all annotations  View by Kinase  Download  Help 

Show Selected	PTM enzyme	PubMed ID	Phosphorylated Protein (Substrate)	Phosphorylation Site	No. of Sentences	Text Evidence
<input type="checkbox"/>	chk1	17339337	p53	Ser-20	1	
		23886938	lats2	Ser-835	1	
		24922656	cdc25c	Ser-216	1	
		22024163	aurora b	Ser-331	1	
		16963448	claspin	Thr-916	1	
		22001744	enos	Ser-1179	3	

Text Evidence Choose a specific section: All 

Back to Views  Download  Layout  Help 

PubMed Information

22001744	2011	Jung-Hyun Park, Wuon-Shik Kim, Jin Yi K..	Free radical biology & medicine	Full Text
----------	------	---	---------------------------------	---------------------------

No.	PTM enzyme	Substrate	Site	Sentence
1	chk1	enos	Ser-1179	7, 8, 14 (Abstract 1)

Text Evidence

Abstract (ABSTRACT 1)

7 Similarly , ectopic expression of small interference RNA for Chk1 or a dominant-negative Chk1 repressed the UV-irradiation stimulatory effect , whereas **wild-type Chk1** increased **basal eNOS-Ser(1179) phosphorylation** .

8 **Purified Chk1** **directly phosphorylated** **eNOS Ser(1179)** in vitro .

14 Overall , this is the first study demonstrating that **Chk1** **directly phosphorylates** **eNOS Ser(1179)** in response to UV irradiation , which is dependent on Hsp90 interaction .

Select/deselect: kinase substrate site phospho keywords

Fig. 4 RLIMS-P “Kinase view” partial results for CHK1 search. The “No. of sentences” column provides a quick link to evidence sentences for the specific annotation

1. Go to the eFIP home page (<http://proteininformationresource.org/efip>).
2. Enter the following protein names in the “Enter Protein Names and Type” query box: CHK1 OR CHEK1 OR “checkpoint kinase-1” and click Submit (Fig. 2a 2). Note that the search can be restricted to retrieve results with CHK1 as a substrate, a kinase, or an interactant (Fig. 2a 3).
3. Select “Substrate View.” After submission, the result page (Fig. 2b) displays the data in a summary view (as a list of entities detected that are grouped by PMID). Similar to RLIMS-P, by

Table 3

Substrates of CHK1 and phosphorylation sites. These are manually validated results from RLIMS-P output. n/a indicates that the phosphorylated site is not described in the RLIMS-P output

CHK1 substrates	Site	PMIDs
AURKB	Ser-331	22024163, 23321637
BLM	Ser-646	20719863
BRCA2	Thr-3387	24627786, 18317453
CDC10	n/a	24006488
CDC2	Tyr-15	24996846, 11479224
CDC25	Ser-287	9744884
	Ser-99	10198041
	n/a	9923681, 11133168, 15272308, 9774107, 10469601, 17912454
CDC25A	Ser-123	12399544, 12759351
	Ser-178, Thr-507	1459997
	Ser-73	12110582
	Ser-75	12759351
	Ser-76	14681206, 20348946, 18480045, 21252624, 20798862
	Thr-504	15272308
	n/a	12110582, 12399544, 20609246, 18414041, 24022480, 19244340, 21851590, 15272308, 19638579, 23272087, 21347609, 9278511, 18480045
CDC25B	Ser-230, Ser-563	17003105
	n/a	9278511, 10713667, 20798862
CDC25C	Ser-216	14681223, 9278511, 10676638, 11027648, 24922656, 15282313, 10557092, 22623962, 23874958, 20700484
	n/a	18272544, 10090724, 9278511, 11479224, 11925443, 15220526, 10681541, 22941630, 20798862, 21347609, 11278490, 10068474, 24038466
CDKI	n/a	20798862
CDKN1A	n/a	21791608

CDKN1C	n/a	21791608
CHK1	Ser-296	23068608, 20053762, 21289283, 24996846
	Ser-317, Ser-345	21851590
	n/a	14681223, 15371427, 23548269, 23593009, 19421147
CK1D	n/a	23861943
CK2	n/a	15225637
CLP1	n/a	22918952
CLSPN	Thr-916	16963448
CRB2	Ser-80	22792081
	Thr-73	22792081
CSNK1D	Ser-328, Ser-331, Thr-397	23861943
	Ser-328, Ser-331, Ser-370, Thr-397	23861943
	Ser-328, Thr-329, Ser-331, Ser-361, Ser-382	23861943
	n/a	23954429
	Ser-1179	22001744
ERRFI1	Ser-251	22505024
FANCD2	n/a	21926477
FANCE	Thr-346, Ser-374	17296736
H2AFX	Ser-345	24913641
H2AX	Thr-16	20639511
KAP1	Ser-473	21851590
LATS2	Ser-408	21118956
	Ser-835	23886938

(continued)

Table 3
(continued)

CHK1 substrates	Site	PMIDs
MAD2	n/a	23454898
MDMX	Ser-367	16511572
p33 (ING1b)	Ser-126	17585055
p50	n/a	22152481
PDS1	n/a	11390356, 17671432
RAD51	n/a	18317453
RAD9	n/a	24376897
RASSF1	Ser-184	24197116
RB1	n/a	17380128
RELA	Ser-612 Thr-505	15970704 17962807
RPAL	n/a	16412704
SETMAR	n/a Ser-495	25024738 22231448
SYK	Ser-295	22585575
TAU	n/a	23550703
TLK1	n/a	12660173
TP53	Ser-695 Ser-20 Ser-23 n/a	24376897, 12955071 15467443, 17339337 23152407 15659650, 11599922, 23272087
TP73	Ser-47	14585975
WEE1	Ser-549	11251070

selecting “Substrate view” the information can be grouped by phosphorylated substrate, so that we can check the PPIs for phosphorylated CHK1.

4. *Select “Kinase view”* to investigate phosphorylation-dependent PPIs for CHK1 substrates. Review results for CHK1 as kinase, and check the information for interactant with its associated text evidence. Figure 5a depicts the text evidence for PMID: [14559997](#). In this particular case, the phosphorylation of CDC25A on Ser-178 and Thr-507 by CHK1 promotes the binding to 14-3-3 proteins. In addition to highlighting kinases, substrates, and sites using the same color scheme as RLIMS-P, interactants are highlighted in orange.

You can also check for information about CHK1 as interactant, using the “Interactant view.”

5. *Download the result table.* Similar to RLIMS-P, eFIP results can be downloaded in CSV format by using the “Download Table” link in the upper left corner of the results table (Fig. 2b). Table 4 provides a summary of results where CHK1 participated in a phosphorylation-dependent PPI either as the phosphorylated substrate or as the interactant. Table 5 provides a summary of phosphorylation-dependent PPIs where CHK1 acts as the kinase.

3.4 Visualization of Phosphorylation and Interaction Events in Cytoscape

eFIP also supports visual exploration of phosphorylation interaction networks using Cytoscape [28], which depicts in one graph a network of kinase–substrate relations, as well as PPI relations, including both the enhancement and inhibition of an interaction. Therefore, the phosphorylation-dependent interactions described in Subheading 3.3 can be displayed in Cytoscape.

Cytoscape View from Text Mining PMID Evidence Page

1. *Go to the eFIP home page.*
2. *Search for PMID:14559997.* Enter the PMID in the search box (Fig. 2a 4) and submit.
3. *Open the “Text Evidence” page.* Click on the “hand” icon in the last column (Fig. 2b 4) to see the text evidence (Fig. 5a).
4. *Click on “See Cytoscape View.”* The link to the Cytoscape view is at the top right of the evidence table (Fig. 5a 5). The Cytoscape view for this example is shown in Fig. 5b. CHK1 phosphorylates CDC25A at two residues. The phosphorylated residues enable interaction with 14-3-3.

Cytoscape View for Multiple Articles (See Note 4)

5. *Go to the eFIP home page.*
6. *Conduct query.* For this case, enter the PMIDs [12676962](#), [17380128](#), and [20639859](#) separated by commas in the search box (Fig. 2a 4) and then submit.

A

Text Evidence (Download Evidence) Back to Search Result See Cytoscape View

5

Chk1 kinase negatively regulates mitotic function of Cdc25A phosphatase through 14-3-3 binding .

Mei-Shya Chen, Christine E Ryan, Helen Piwnica-Worms

Molecular and cellular biology, 2003

14559997

Choose a specific sub-section: All 1

Each row of the table can be read as: Protein **Kinase** phosphorylation of **Substrate** at **Site** Impacts association/dissociation of **Substrate** with **Interactant**.

No.	Protein Kinase	Substrate	Site	Impact	Interactant	Section Type	Sentence
1	chk1	cdc25a	Thr-507, Ser-178	enables association	14-3-3	Abstract	4

2

Gene Normalization + ?			
Protein	Name	UniProtKB AC	Annotation No.
Kinase	chk1	O14757/CHK1_HUMAN B4DT73/B4DT73_HUMAN	1
Substrate	cdc25a	P30304/MPIP1_HUMAN	1
Interactant	14-3-3	P27348/1433T_HUMAN	1

3

Text Evidence

Abstract

4

...

4 **Phosphorylation** of **Cdc25A** on **serine 178** and **threonine 507** **facilitates** **14-3-3** **binding** , and **Chk1** **phosphorylates** both residues in vitro .

...

Select/deselect: kinase substrate site interactant phospho keywords PPI keywords impact keywords

B

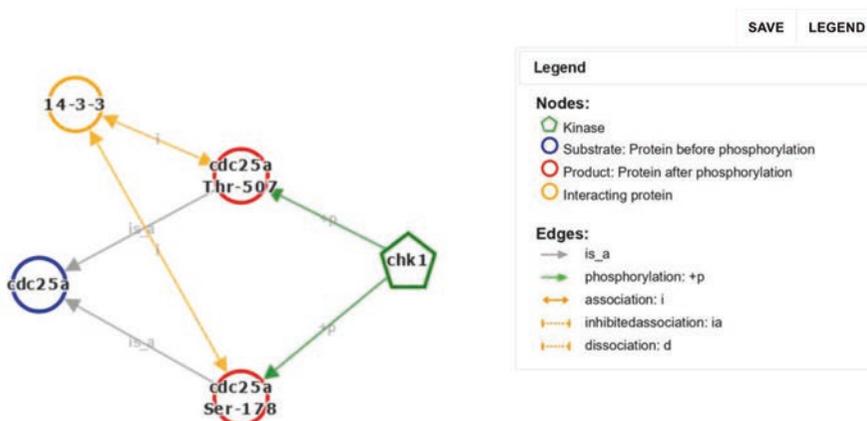


Fig. 5 Analysis of CHK1 eFIP text evidence for PMID:14559997. (a) The results can be filtered by section of the article (1). The table with extracted results shows the substrate, kinase (if known), phosphorylation site, the impact (inhibit, enhance, unknown association/dissociation), the interactant, the section of the article where the information comes from and the sentence number in that section (2). The gene normalization table (3)

Table 4
Summary of eFIP results for phosphorylation-dependent PPI involving CHK1

Substrate	PMIDs	Kinase	Site	Impact	PPI	Interactant	
CHK1	12676962	CHK1	Ser-345	Enables	Association	14-3-3	
	12415000		Ser-345	Enables	Association	RAD24 (14-3-3 homolog)	
	15585577				Enables	Association	RAD25 (14-3-3 homolog)
	20639859		Ser-296	Enables	Association	CDC25A	
	12676962		Ser-345	Enables	Association	Chromatin	
	16360315				Enables	Dissociation	Chromatin
	23593009		Thr-125	Inhibits	Association	RAD9	
	23593009		Thr-143	Enables	Association	RAD9	
CRB2	22792081			Increases	Association	CHK1	
CLSPN	15707391		Thr-916, Ser-945	Enables	Association	CHK1	
	22792081			Increases	Association	CHK1	
	12766152			Unknown	Association	CHK1	
	12545175			Unknown	Association	CHK1	

Table 5
Summary of eFIP results for CHK1 substrates with phosphorylation-dependent PPIs

Substrate	PMIDs	Kinase	Site	Impact	PPI	Interactant
BRCA2	18317453	CHK1		Enables	Association	RAD51
CDC25	23166842	CHK1		Increases	Association	SCFBETATrCP
	22806395			Enables	Association	RAD24 (14-3-3 homolog)
CDC25A	14559997	CHK1	Thr-507, Ser-178	Enables	Association	14-3-3
	15272308		Thr-504	Enables	Association	14-3-3
	15272308		Thr-504	Inhibits	Association	Cdk1-cyclin A Cdk1-cyclin B Cdk2-cyclin E
CDC25B	20798862	CHK1	Ser-309, Ser-323	Enables	Association	14-3-3
CDC25C	23874958	CHK1	Ser-216	Increases	Association	14-3-3beta
	20798862		Ser-216	Enables	Association	14-3-3
RB1	17380128	CHK1	Ser-612	Enables	Association	E2F1
RAD51	18317453	CHK1		Enables	Association	BRCA2

←
Fig. 5 (continued) shows possible UniProt identifiers for the kinase, substrate, and interactant mentioned in the table. The text evidence panel (4) contains the evidence sentences with section title and sentence numbers.
(b) The Cytoscape representation of the relations extracted in the article

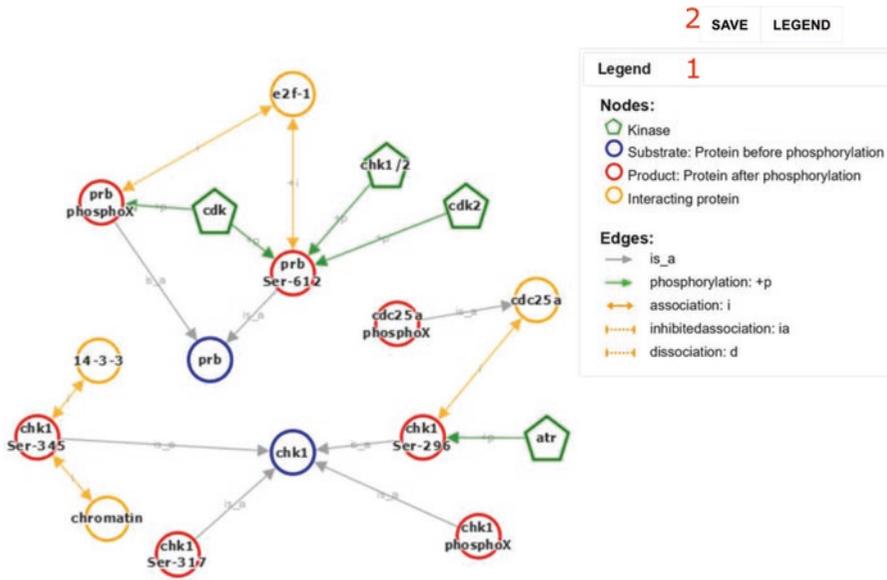


Fig. 6 The Cytoscape representation for the phosphorylation and PPI events extracted from articles with PMID [12676962](#), [17380128](#), and [20639859](#). Events extracted are illustrated with nodes and edges; see legend for details (1). The graph can be exported (2)

7. Open “Cytoscape View.” The link to Cytoscape is on the top left of the result table (Fig. 2b 5). The Cytoscape view for this example is shown in Fig. 6.

4 Notes

1. The “Login” link is located in the upper-right corner of the web page. When you click it, it will ask you to either enter your credentials or sign up. Select sign up and complete the information needed. After the registration, an automatic email will be sent out to explain details on how to log into RLIMS-P or eFIP.
2. For our CHK1 query, we obtained 1266 articles with potential protein phosphorylation mentions, with 278 kinase mentions, 854 substrate mentions, and 245 site mentions as of 05/27/2016. Note that if you query PubMed instead of RLIMS-P with the same query, it retrieves many more articles, 2781 as of 05/27/2016, many of which are not relevant to CHK1 phosphorylation at all. In PubMed, finding the subset where CHK1 is phosphorylated could then only be achieved by manual inspection, whereas in RLIMS-P, selecting the appropriate view will enable quick access to the most relevant set.
3. The text mining results should not be assumed to be completely correct. There is a possibility of encountering false positive results or of missing relevant data. The different tools

provide their own metrics of performance, and it is important to be aware of them when using the tools. In addition, one should consider reviewing the substrate names thoroughly, as textual variants of CHK1 presently appear as separate substrates. We are currently working on improving the consistency and grouping of the substrate and kinase names.

4. The Cytoscape view provides an overview of the text mining results in graphical format. However, if the output includes multiple articles the number of nodes and edges may become overwhelming. The current version of Cytoscape used in eFIP does not allow hiding or displaying selected nodes. To see only a selected subset, one could either retrieve data for selected PMIDs or, alternatively, download a desktop version of Cytoscape to read the saved XGMML-beta (Cytoscape compatible) file (Fig. 6 2).

Acknowledgments

This work was supported by grants from the National Institutes of Health: R01GM080646 and U01HG008390.

References

1. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 43(Database issue):D512–D520. doi:[10.1093/nar/gku1267](https://doi.org/10.1093/nar/gku1267)
2. Steelman LS, Martelli AM, Cocco L, Libra M, Nicoletti F, Abrams SL, McCubrey JA (2016) The therapeutic potential of mTOR inhibitors in breast cancer. *Br J Clin Pharmacol*. doi:[10.1111/bcp.12958](https://doi.org/10.1111/bcp.12958)
3. Yamaoka K (2016) Janus kinase inhibitors for rheumatoid arthritis. *Curr Opin Chem Biol* 32:29–33. doi:[10.1016/j.cbpa.2016.03.006](https://doi.org/10.1016/j.cbpa.2016.03.006)
4. Wang Y, Ma H (2015) Protein kinase profiling assays: a technology review. *Drug Discov Today Technol* 18:1–8. doi:[10.1016/j.ddtec.2015.10.007](https://doi.org/10.1016/j.ddtec.2015.10.007)
5. de Oliveira PS, Ferraz FA, Pena DA, Pramio DT, Morais FA, Schechtman D (2016) Revisiting protein kinase-substrate interactions: toward therapeutic development. *Sci Signal* 9(420):re3. doi:[10.1126/scisignal.aad4016](https://doi.org/10.1126/scisignal.aad4016)
6. Ross KE, Arighi CN, Ren J, Huang H, Wu CH (2013) Construction of protein phosphorylation networks by data mining, text mining and ontology integration: analysis of the spindle checkpoint. *Database (Oxford)* 2013:bat038. doi:[10.1093/database/bat038](https://doi.org/10.1093/database/bat038)
7. Fleuren WW, Alkema W (2015) Application of text mining in the biomedical domain. *Methods* 74:97–106. doi:[10.1016/j.ymeth.2015.01.015](https://doi.org/10.1016/j.ymeth.2015.01.015)
8. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B (2013) Biomedical text mining and its applications in cancer research. *J Biomed Inform* 46(2):200–211. doi:[10.1016/j.jbi.2012.10.007](https://doi.org/10.1016/j.jbi.2012.10.007)
9. Huang CC, Lu Z (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief Bioinform* 17(1):132–144
10. Hu Z-Z, Mani I, Hermoso V, Liu H, Wu CH (2004) iProLINK: an integrated protein resource for literature mining. *Comput Biol Chem* 28(5–6):409–416. doi:[10.1016/j.compbiolchem.2004.09.010](https://doi.org/10.1016/j.compbiolchem.2004.09.010)
11. Peng Y, Torii M, Wu CH, Vijay-Shanker K (2014) A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems. *BMC Bioinform* 15:285. doi:[10.1186/1471-2105-15-285](https://doi.org/10.1186/1471-2105-15-285)
12. Peng Y, Tudor C, Torii M, Wu C, Vijay-Shanker K (2012) iSimp: a sentence simplification system for biomedical text. *International Conference on Bioinformatics and Biomedicine (BIBM2012)*:211–216

13. Ding R, Arighi CN, Lee JY, Wu CH, Vijay-Shanker K (2015) pGenN, a gene normalization tool for plant genes and proteins in scientific literature. *PLoS One* 10(8):e0135305. doi:[10.1371/journal.pone.0135305](https://doi.org/10.1371/journal.pone.0135305)
14. Tudor CO, Ross KE, Li G, Vijay-Shanker K, Wu CH, Arighi CN (2015) Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. *Database* 2015:bav020
15. Tudor CO, Arighi CN, Wang Q, Wu CH, Vijay-Shanker K (2012) The eFIP system for text mining of protein interaction networks of phosphorylated proteins. *Database* 2012:bas044
16. Tudor CO, Schmidt CJ, Vijay-Shanker K (2010) eGIFT: mining gene information from the literature. *BMC Bioinf* 11:418. doi:[10.1186/1471-2105-11-418](https://doi.org/10.1186/1471-2105-11-418) 1471-2105-11-418 [pii]
17. Torii M, Arighi CN, Li G, Wang Q, Wu CH, Vijay-Shanker K (2015) RLIMS-P 2.0: a generalizable rule-based information extraction system for literature mining of protein phosphorylation information. *IEEE/ACM Trans Comput Biol Bioinform* 12(1):17–29
18. Torii M, Li G, Li Z, Oughtred R, Diella F, Celen I, Arighi CN, Huang H, Vijay-Shanker K, Wu CH (2014) RLIMS-P: an online text-mining tool for literature-based extraction of protein phosphorylation information. *Database* 2014:bau081
19. Li G, Ross KE, Arighi CN, Peng Y, Wu CH, Vijay-Shanker K (2015) miRTex: a text mining system for miRNA-gene relation extraction. *PLoS Comput Biol* 11(9)
20. Xu N, Lao Y, Zhang Y, Gillespie DA (2012) Akt: a double-edged sword in cell proliferation and genome stability. *J Oncol* 2012:951724. doi:[10.1155/2012/951724](https://doi.org/10.1155/2012/951724)
21. Wei CH, Kao HY, Lu Z (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed Res Int* 918710(10):25
22. Natale DA, Arighi CN, Blake JA, Bult CJ, Christie KR, Cowart J, D'Eustachio P, Diehl AD, Drabkin HJ, Helfer O, Huang H, Masci AM, Ren J, Roberts NV, Ross K, Ruttenberg A, Shamovsky V, Smith B, Yerramalla MS, Zhang J, AlJanahi A, Celen I, Gan C, Lv M, Schuster-Lezell E, Wu CH (2014) Protein ontology: a controlled structured network of protein entities. *Nucleic Acids Res* 42(Database issue):21
23. Sanchez Y, Wong C, Thoma RS, Richman R, Wu Z, Piwnica-Worms H, Elledge SJ (1997) Conservation of the Chk1 checkpoint pathway in mammals: linkage of DNA damage to Cdk regulation through Cdc25. *Science (New York, NY)* 277(5331):1497–1501
24. McNeely S, Beckmann R, Bence Lin AK (2014) CHEK again: revisiting the development of CHK1 inhibitors for cancer therapy. *Pharmacol Ther* 142(1):1–10. doi:[10.1016/j.pharmthera.2013.10.005](https://doi.org/10.1016/j.pharmthera.2013.10.005)
25. Goto H, Kasahara K, Inagaki M (2015) Novel insights into Chk1 regulation by phosphorylation. *Cell Struct Funct* 40(1):43–50. doi:[10.1247/csf.14017](https://doi.org/10.1247/csf.14017)
26. UniProt C (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43(Database issue):D204–D212. doi:[10.1093/nar/gku989](https://doi.org/10.1093/nar/gku989)
27. Coordinators NR (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44(D1):D7–D19
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11):2498–2504

Part III

Protein Network Bioinformatics

Chapter 11

Functional Interaction Network Construction and Analysis for Disease Discovery

Guanming Wu and Robin Haw

Abstract

Network-based approaches project seemingly unrelated genes or proteins onto a large-scale network context, therefore providing a holistic visualization and analysis platform for genomic data generated from high-throughput experiments, reducing the dimensionality of data via using network modules and increasing the statistic analysis power. Based on the Reactome database, the most popular and comprehensive open-source biological pathway knowledgebase, we have developed a highly reliable protein functional interaction network covering around 60 % of total human genes and an app called ReactomeFIViz for Cytoscape, the most popular biological network visualization and analysis platform. In this chapter, we describe the detailed procedures on how this functional interaction network is constructed by integrating multiple external data sources, extracting functional interactions from human curated pathway databases, building a machine learning classifier called a Naïve Bayesian Classifier, predicting interactions based on the trained Naïve Bayesian Classifier, and finally constructing the functional interaction database. We also provide an example on how to use ReactomeFIViz for performing network-based data analysis for a list of genes.

Key words Functional interaction, Biological network, Biological pathway, Reactome, Network-based analysis, ReactomeFIViz, Cytoscape, Naïve Bayesian Classifier, Java, MySQL

1 Introduction

Large-scale data sets are routinely generated in current biological studies using high-throughput techniques in order to understand disease mechanisms and develop better personalized precision therapies for patients. However, these data sets are usually gene or protein based. To understand the relationships among interesting genes or proteins, researchers usually have to project them onto biological pathways and network contexts so that these seemingly scattered genes or proteins can be visualized and studied together as groups.

Proteins and other gene products (e.g., miRNAs, lncRNAs) interact with each other and form a huge interaction network inside the cell. High-throughput experiments, such as Yeast 2-Hybrid

(Y2D) [1] and mass spectrometry coupled with co-immunoprecipitation (MS CoIP) [2], have been used to generate large-scale protein–protein interaction data sets for many species, including human, mouse, fly, worm, and yeast. However, usually interaction data sets generated by these experiments have high false positive rates and physical interactions detected by these methods may not necessary play functional roles inside cells.

Based on highly reliable, expert curated pathways in the Reactome knowledgebase [3], the most popular and comprehensive open-source biological pathway database, we have developed a software pipeline to construct a human protein/gene functional interaction (FI) network covering around 60 % of total human genes. Figure 1 shows the overview of procedures and data sources to construct this FI network. We used protein pairwise relationships from protein physical interactions in human, mouse, fly, worm, and yeast, gene co-expression, Gene Ontology (GO) annotation, and protein domain interactions as features to train a Naïve Bayes Classifier (NBC) based on FIs extracted from Reactome. NBC is a simple machine learning technique, assuming independence

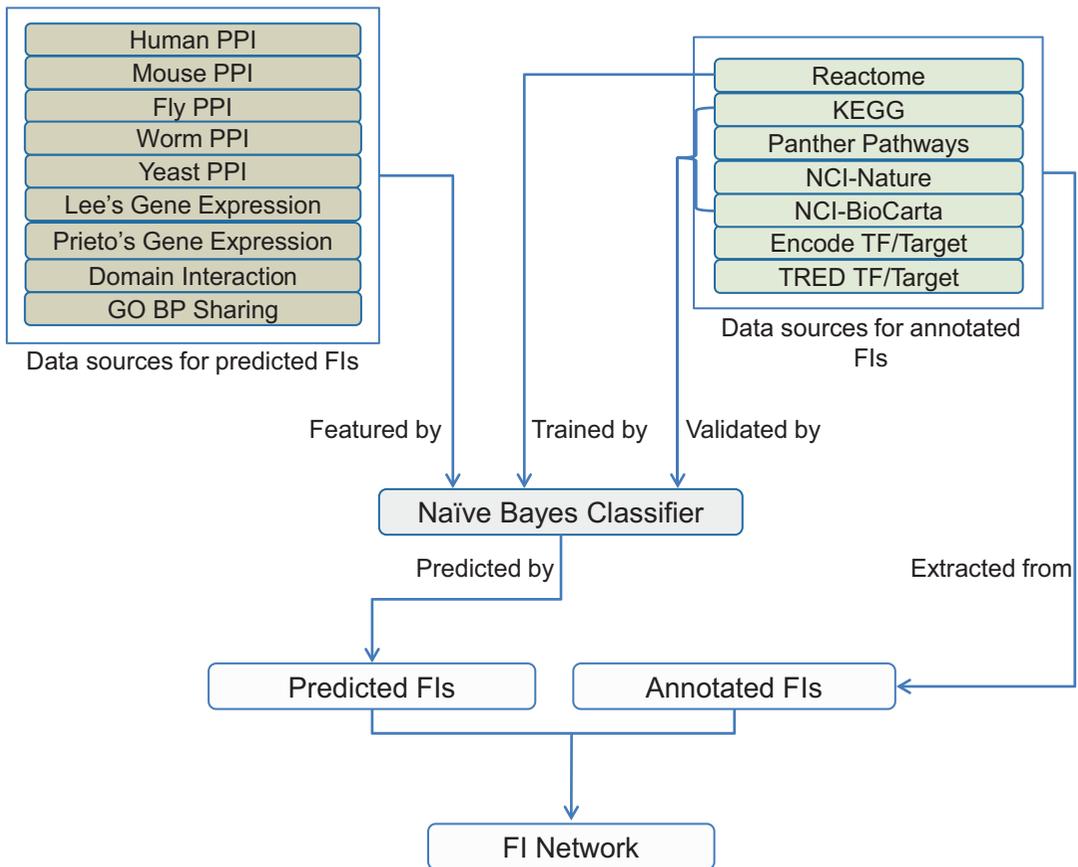


Fig. 1 An overview of the procedures used to construct the Reactome functional interaction network

among features. The trained NBC then was validated by independent FIs extracted from other non-Reactome pathway databases. The final FI network contains two types of FIs: annotated FIs extracted from curated pathway data and predicted FIs by the NBC. For gene regulatory network analysis, we have also included interactions between transcription factors and their targets from the ENCODE project [4] and the TRED database [5].

For users to perform network-based data analysis using the Reactome FI network, we have also developed an app called ReactomeFIViz [6] for Cytoscape [7], the most popular biological network visualization and analysis platform. Users of our app can construct an FI subnetwork for a list of genes, perform network clustering to find network modules, annotate the subnetwork and modules, and then perform survival analysis for network modules.

In this chapter, we describe the procedures to construct the Reactome FI network and briefly introduce ReactomeFIViz. For more information, the reader is encouraged to refer to our other published materials [6, 8].

2 Materials

1. To construct the Reactome FI network, the reader is required to have some programming experience in Java and experience on using Eclipse, the most popular Java IDE (integrated development environment). Eclipse can be downloaded from <http://www.eclipse.org> and the source code for the FI construction can be downloaded from <http://reactomedev.oicr.on.ca/download/tools/caBIG/FINetworkBuild.zip>.
2. You will also need to install mysql database locally after downloading it from <http://www.mysql.org>.
3. To draw ROC (receiver operating characteristic) curves for checking the performance of the Naïve Bayes Classifier, you need to install R. You can download the latest version of R from <https://www.r-project.org>.
4. It is necessary to download many data sources from the Internet to construct the Reactome FI network. We will give URLs in Section 3 “Methods” when we describe detailed procedures.
5. To use ReactomeFIViz, the reader should install the latest version of Cytoscape from <http://www.cytoscape.org>. In order to use Cytoscape, the following hardware and software requirements are suggested:
6. Hardware Requirements: A 2 Ghz or higher dual core or quad core CPU, 4GB or more physical memory, 512 MB video memory, 1 GB or more available hard drive space, a display that supports 1024 × 768 or higher resolution, and a high-speed Internet connection.

7. Software Requirements: To use the latest version of Cytoscape (version 3.3.0 in January, 2016), the reader is required to install Java 8 first. Java 8 is supported under Windows, MacOS, and Linux, and can be downloaded from <http://www.java.com>.

3 Methods

3.1 Setting Up Eclipse and Loading the FINetworkBuild Project

1. The project used to build the FI network is programmed in Java, and Eclipse is used for Java programming. After downloading Eclipse from <http://www.eclipse.org> and the Java source code for the project from <http://reactomedev.oicr.on.ca/download/tools/caBIG/FINetworkBuild.zip>, use menu File/Import to get the Project dialog and choose “Existing Projects into Workspace” (Fig. 2a), and then follow the step-by-step procedures to create a Java project in Eclipse (Fig. 2b).
2. The configuration file (File name: configuration.prop) in the resources folder is used to configure files and parameters used by the project. Open this file in Eclipse and make sure the following three values are configured to what you want (*see Note 1* for suggested values): YEAR, RESULT_DIR, DATA_SET_DIR.

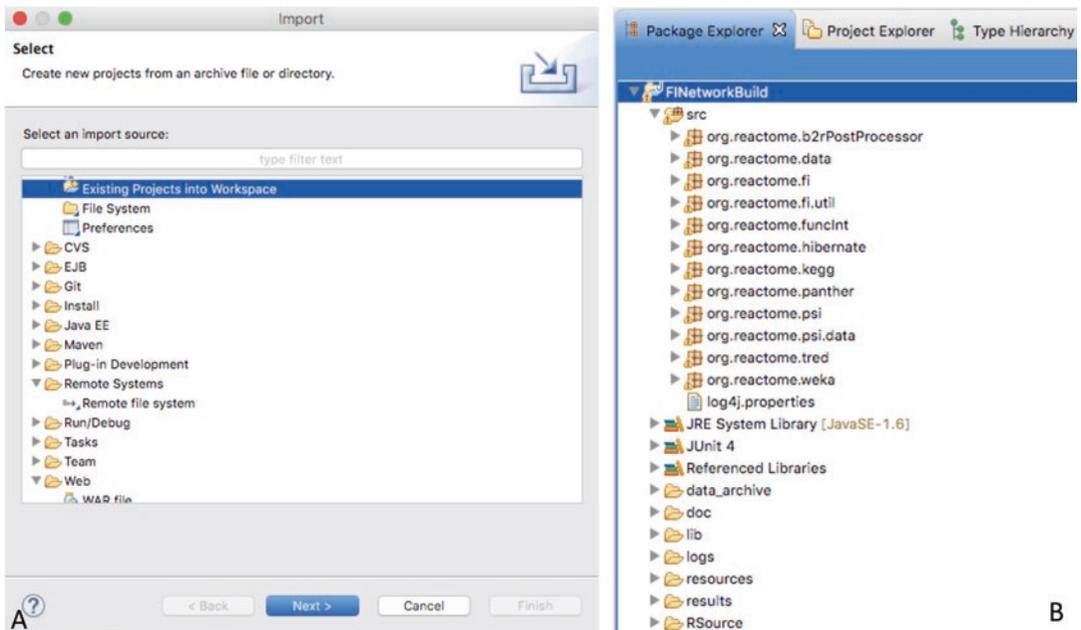


Fig. 2 Import the Java source code to build a project for constructing the Reactome FI network. (a) (left): the dialog for importing the Java source code into a project in Eclipse. (b) (right): the package explorer view of the imported Java project for constructing the FI network

3. We use log4j for logging output from program runs. The configuration for log4j is in file `log4j.properties` in the resource folder. Since a large amount of output will be generated during some methods' running, it is recommended that you use a file to keep the log output (*see Note 2* for some discussion on how to configure log4j).

3.2 Data Sources for ID Mapping

The construction of the Reactome FI network relies on many data sources, which may use different identifiers for genes and proteins. We use identifiers from UniProt [9] as the standard and PIR [10] for ID mapping.

1. UniProt is the authoritative protein knowledgebase used in FI network construction. We map all proteins and genes to UniProt accession numbers for normalization based on protein amino acid sequences to remove duplications. In this case, you will need two files from the directory `current_release/knowledgebase/taxonomic_divisions` in its ftp site, [ftp.uniprot.org](ftp://ftp.uniprot.org): `uniprot_sprot_human.data.gz` and `uniprot_trembl_huma.dat.gz`.

For normalizing features used by the NBC, you also need another file from UniProt for protein isoform sequences: `uniprot_sprot_varsplic.fasta`. This file should be downloaded from the UniProt web site: <http://www.uniprot.org/downloads>, searching for Isoform sequences. The downloaded file should be named as is.

All three files downloaded from UniProt should be placed in the same folder. The folder name should be configured in the configuration file (`configuration.prop`) for property, `UNIPROT_DIR` (*see Note 3* for a suggested name).

2. We also use a mapping file to map Entrez gene ids to UniProt accession numbers. This file can be downloaded from the PIR ftp site <ftp.pir.georgetown.edu/databases/>: after logging in, go to `idmapping/mapping_by_sp`, and download file `h_sapiens.tb`. Remember where you place your file, and modify these two properties in the configuration file: `IPOCLASS_HUMAN_FILE` and `ENTREZ_TO_UNIPROT_MAP_FILE_NAME`. The second property will be used by the program later on.

3.3 Data Sources for Annotated FIs

Annotated FIs are FIs extracted from human curated pathways. In addition to Reactome, we have also extracted FIs from KEGG [11], NCI-PID [12], and Panther Pathways [13]. Pathways in other non-Reactome databases are imported into the Reactome curator tool project first for easy checking and then dumped into an enhanced Reactome database locally. In addition to pathways, we also import two interaction data sets for transcription factors and their targets from TRED [5] and ENCODE [4] as annotated FIs.

1. Reactome is used as the foundation for the FI network building. We use a slice of the Reactome central database to import data from other sources. You can find this database in the results folder after getting the project source. In addition to this slice database, we also need a snapshot of the original central database. For the latest version, send a help email to help@reactome.org.

Unzip these two database dump files and import them into your local mysql after logging into it. For example, for loading the slice database from the mysql dump file, test_slice_55.sql, use the following commands:

```
create database reactome_55_plus_i;
```

```
use reactome_55_plus_i;
```

```
source test_slice_55.sql (Refer to Note 4 about loading the dump file into mysql).
```

After loading these two databases into your local mysql, modify the following four properties in the configuration file with values you are using for your databases: REACTOME_SOURCE_DB_NAME, DB_USER, DB_PWD, REACTOME_GK_CENTRAL_DB_NAME.

We will modify the loaded slice database, called reactome_55_plus_i in the aforementioned example, for the FI network construction. For database transaction protection, we need to change the table type in the database from MyISAM to InnoDB. To make this change, run the JUnit method changeMyISAMToInnoDB() in class org.reactome.data.ReactomeDatabaseModifier.

The slice database contains only some of the human proteins in UniProt. Copy other human proteins from the snapshot of the curated database by running the Java method, copyHumanReferenceGeneProducts().

Finally, we need to modify the database schema by adding a new attribute, dataSource, and two new classes, Interaction and Targeted Interaction by running the following command in the mysql terminal:

```
source {absolute_path_to}/SchemaModification.sql.
```

You should find a copy of SchemaModification.sql in the resources folder. Refer to **Note 5** on how to check if your database schema has been updated.

2. KEGG has many good pathway diagrams and disease pathways. We download needed files from KEGG's ftp site: [ftp.bioinformatics.jp](ftp://ftp.bioinformatics.jp). However, you will need to have a license first in order to access its ftp site. The following files are needed: kegg/xml/kgml/non-metabolic/organisms/hsa.tar.gz, kegg/pathway/pathway.list, /kegg/pathway/map_title.tab, kegg/pathway/organisms/has.tar.gz, kegg/genes/links/genes_uniprot.list.gz. Unzip these files and then extract a much smaller mapping file for human proteins using the following command (mac and linux only):

grep hsa: genes_uniprot.list > hsa_genes_uniprot.list.

Specify values in the configuration file for the properties related to KEGG based on the directory in which you have placed your downloaded files (*see Note 3* for more information): KEGG_DIR, KEGG_HSA_KGML_DIR, KEGG_CONVERTED_FILE, KEGG_ID_TO_UNIPROT_MAP_FILE.

3. NCI-PID is a database for cancer pathways, composed of two sources: pathways curated by NCI-PID curators and pathways imported from BioCarta and Reactome. For constructing the FI network, we will import NCI-PID curated pathways and pathways from BioCarta using their BioPAX Level 2 export, which can be downloaded from <http://pid.nci.nih.gov/download.shtml>. After unzipping, you should find two files: NCI-Nature_Curated.bp2.owl, BioCarta.bp2.owl. (*see Note 6* about the NCI-PID database).

For NCI-PID, the following parameters need to be specified in the configuration file based on the directory you have used (refer to **Note 3**): NATURE_PID_DIR, NATURE_PID_CURATED, NATURE_PID_CURATED_CONVERTED, NATURE_PID_BIOCARTA, NATURE_PID_BIOCARTA_CONVERTED.

4. Pathways in the Panther database are imported based on the SBML format. They are downloaded from Panther's ftp site: <ftp.pantherdb.org>. Two files are needed in directory pathway/current_release: SBML_{version}.zip and SequenceAssociationPathway{version}.txt ({version} should be replaced by an actual version number, e.g., 3.0.1), and these properties should be set in the configuration file based on the directory you use (refer to **Note 3**): PANTHER_DIR, PANTHER_FILES_DIR, PANTHER_MAPPING_FILE, PANTHER_CONVERTED_FILE.
5. TRED provides functional interactions between transcription factors (TFs) and their targets. Two types of TF/target interactions are available in TRED: human curated and computational predicted. For constructing the FI network, we extract human curated ones only using the Hibernate API (<http://hibernate.org/>) after installing the TRED database locally in mysql. You can get a mysql dump file from data_archive/tred. You also need to configure the Hibernate configuration file (resources/TREDHibernate.cfg.xml) based on your database and change these two properties (Refer to **Note 3**): TRED_DIR, TRED_CONVERTED_FILE.
6. ENCODE TF/target interactions were generated by the Gerstein group in Yale (<http://encodenets.gersteinlab.org>), originally

published in Nature [4]. For constructing the FI network, we have kept a file containing these interactions in `data_archive/encode/tf-targets.txt`. To use this file, specify these properties in the configuration file according to names of the directory and the file (refer to **Note 3**): `ENCODE_DIR`, `ENCODE_TFF_FILE`, `ENCODE_TFF_CONVERTED_FILE`. (Refer to **Note 7** for more information on using this data set.)

3.4 Data Sources for Predicted FIs

We use multiple features to predict functional interactions between two proteins, including physical interactions downloaded from iRefIndex [14] for human, mouse, fly, worm, and yeast, gene co-expression from two sources [15, 16], GO [17] annotation, and protein domain–domain interactions. Protein interactions from non-human species are mapped to human using Ensembl-Compara [18].

1. To download files from Ensembl-Compara, go to its download page: <http://www.ensembl.org/info/data/ftp/index.html>, choose Comparative/MySQL to go to its ftp site. After logging into its ftp site, choose `ensembl_compara_xx` (xx for release number), and then download `seq_member.txt.gz`, `family.txt.gz`, `family_member.txt.gz`, and `ensembl_compara_xx.sql.gz`. Unzip all files.

Log into your mysql database, create an empty database named `ensembl_compara_xx` (xx for release number), and load the database schema with command: `source ensembl_compara_xx.sql`. Log out from mysql and load downloaded data files with the following command:

```
mysqlimport -u{mysql_db_user} -p --local ensembl_compara_xx family.txt family_member.txt seq_member.txt
```

(Refer to Note 8 for how to make sure you have loaded content correctly.)

Assign correct values to the following properties in the configuration file based on values you are using for your `ensembl_compara` database: `ENSEMBL_DIR`, `ENSEMBL_COMPARA_DATABASE`, `ENSEMBL_PROTEIN_FAMILIES` (This file will be autogenerated in Subheading 3.5).

2. All protein–protein interactions used are downloaded from iRefIndex [14], which collects interaction data from several sources and then normalize them based on amino acid sequences and UniProt accession numbers. These interactions are provided in the PSIMI-TAB format and can be downloaded from iRefIndex’s ftp site via <http://irefindex.org/wiki/index.php?title=iRefIndex:6239.mitab.04072015.txt.zip> (worm), [7227.mitab.04072015.txt.zip](http://irefindex.org/wiki/index.php?title=iRefIndex:7227.mitab.04072015.txt.zip) (fly), [9606.mitab.04072015.txt.zip](http://irefindex.org/wiki/index.php?title=iRefIndex:9606.mitab.04072015.txt.zip) (human), [10090.mitab.04072015.txt.zip](http://irefindex.org/wiki/index.php?title=iRefIndex:10090.mitab.04072015.txt.zip) (mouse), [559292.mitab.04072015.txt.zip](http://irefindex.org/wiki/index.php?title=iRefIndex:559292.mitab.04072015.txt.zip) (yeast S288c). Unzip these files, and update the properties with names starting with `IREFINDEX` in the configuration file.

3. We use two gene co-expression data sets for NBC training and prediction: Lee's gene expression [15] and Prieto's gene expression [16]. You can get these two data files from `data_archive/LeeGeneExp` and `data_archive/PrietoGeneExp` (*see Note 9* about how we obtained these two data files), and make sure these two properties are correct in the configuration file: `LEE_GENE_EXP_FILE_SOURCE`, and `PRIETO_PAIRS_FILE`.
4. We use GO biological process (BP) annotation to check if a pair of genes has shared GO BP terms. You will need two files from GO: `gene_association.goa_human` from <http://www.geneontology.org/GO.downloads.annotations.shtml>, and `GO.terms_and_ids.txt` from http://www.geneontology.org/doc/GO.terms_and_ids. Make sure the file names are as described here and the `GO_DIR` value is correct in the configuration file.
5. We use domain–domain interactions from the pFam database [19]. To download domain files, log into pFam ftp site: <ftp.ebi.ac.uk>, go to the latest release folder (e.g., `/pub/databases/Pfam/releases/Pfam29.0` in December, 2015), get these two files: `pfamA.txt.gz` and `pfamA_interactions.txt.gz`, and then unzip them. Make sure the value for property `PFAM_DIR_NAME` correct in the configuration file.

3.5 Construction of the Reactome FI Network

We have developed a software pipeline to construct the FI network. Usually you should simply run methods in Eclipse in the following order after all source files and databases have been downloaded, set up and configured correctly as described previously. The following methods are collected in class `org.reactome.fi.FINetworkBuilder` (refer to **Note 2** about logging).

1. `prepareMappingFile`: After this method runs, four files should be generated: `Uni2Pfam.txt`, `SwissProtACIDMap.txt`, `ACIDMap.txt` in the `UniProt` directory, and `ENTREZ_TO_UNIPROT_MAP_FILE_NAME` (*see* the actual file names in your configuration file) in the `iproclass` directory.
2. `convertPathwayDBs` (refer to **Note 10**): Convert Pathways in KEGG, NCI-PID, and Panther, and TF/Target interactions in TRED and ENCODE into their own respective curator tool project files with extension names `.rtpj`. You may need to open these converted project files in the Reactome curator tool to see how they look and make sure they are correct.
3. `dumpPathwayDBs` (refer to **Note 11**): Dump the converted curator tool projects into the extended Reactome database created at **step 1** in Subheading **3.3**.
4. `dumpPathwayFIs`: Extract annotated FIs from individual pathway sources.

5. prepareNBCFeatures: Check individual features used for training the NBC and generate necessary files for training.
6. trainNBC (refer to **Note 12**): Before running this method, make sure you have assigned correct values to these two properties in the configuration file: ROC_CURVE_FILE and BP_DOMAIN_SHARED_PAIRS.

After finishing the running of this method, draw an ROC (recover operating characteristic) curve to estimate the performance of the trained NBC by running an R script in the RSource folder, ROCcurveDrawing.R, after modifying the fileName variable in the code. You should get an ROC curve similar to Fig. 3.

To calculate AUC (area under curve) of the drawn ROC, you need to install the ROC package from this web page: <http://www.bioconductor.org/packages/release/bioc/html/ROC.html>, and run a command like this in the R console: calculate.AUC(rocData). The result should be greater than 85 % usually.

7. predictFIs: Before running this method, set the cutoff value in the configuration file (usually it should be 0.50) using property CUT_OFF_VALUE and the file name for the predicted FIs, PREDICTED_FI_FILE.

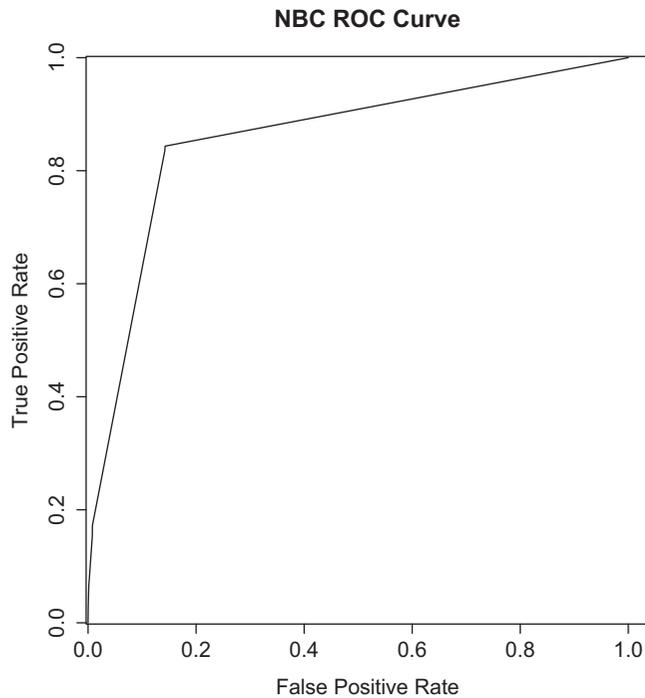


Fig. 3 An ROC curve drawn based on data points generated from method trainNBC

8. buildFIDb: Before running the method, create an empty mysql database named as “FI_yyyy” (yyyy should be the year when the FI network is constructed), and make sure the following values in the Hibernate configuration file in the resources folder, funcIntHibernate.cfg.xml, are correct:

```
<property name="connection.url">jdbc:mysql://localhost:3306/FI_YYYY</property>
<property name="connection.username">{mysql_user_name}</property>
<property name="connection.password">{mysql_user_password}</property>
```

3.6 Installing Cytoscape

In previous sections, we described the procedures to construct the Reactome FI network stored in a mysql database (*see Note 13* for the total time to construct the Reactome FI network). To allow researchers to perform network-based data analysis using this network easily, we have developed a Cytoscape app called ReactomeFIViz [6]. Cytoscape [7] is the most popular open-source network visualization and analysis platform. ReactomeFIViz is an application or “app” that extends the functionality of Cytoscape to explore Reactome pathways and search for disease-related pathways and network patterns using the Reactome FI network. In the following sections, we describe how to use ReactomeFIViz starting with installation of Cytoscape.

To install Cytoscape, follow these two steps:

1. Use your web browser and load the Cytoscape web site (<http://www.cytoscape.org/>), select the “Download” button, and follow the instruction for user registration and installing the software. Refer to **Note 14** for alternative Cytoscape installation methods.
2. Once installed, launch the Cytoscape application. Refer to **Note 15** for more information about how to launch Cytoscape.

3.7 Installing ReactomeFIViz

There are two ways to install ReactomeFIViz: directly from within the Cytoscape software (described in the following) or from the Cytoscape App Store (<http://apps.cytoscape.org>) (*see Note 16* for this method).

1. Launch the Cytoscape software.
2. Go to the “Apps” drop-down menu and select the “App Manager...” feature.
3. In the “Search:” text box, type “ReactomeFI” (remove the quotation marks).
4. Select the “ReactomeFIPlugin” App (ReactomeFIViz was called ReactomeFIPlugin previously).
5. Click on the “Install” button.

3.8 Using ReactomeFIViz to Analyze a Gene List

ReactomeFIViz accesses the Reactome FI network allowing the user to construct an FI subnetwork based on a set of genes, query the FI data source for the underlying evidence for the interaction,

build and analyze network modules of highly connected sets of genes, perform functional enrichment analysis to annotate the modules with pathway or GO annotations, and overlay a variety of information sources such as NCI Cancer Gene Index or GeneCard annotations.

In this section, we use a simple gene list derived from a glioblastoma multiforme (GBM) study [20] to demonstrate how to analyze a list of genes using the Gene Set/Mutation Analysis feature of ReactomeFIViz. For further details about the different supported file formats, *see* **Note 17**. For other features provided in ReactomeFIViz, consult the tutorial in <http://wiki.reactome.org/index.php/ReactomeFIViz>.

1. Launch the Cytoscape software.
2. Go to “Apps” in the drop-down menu, select “Reactome FI,” and then click the “Gene Set/Mutational Analysis” feature. A pop-up window will appear allowing upload of the gene list and configuration of the Gene Set/Mutation Analysis feature.
3. Choose a Reactome FI Network Version from the listed three versions. For this worked example, the most recent (2014) was selected (*see* **Note 18**).
4. Under File Parameters, choose a file containing genes you want to use to construct a functional interaction network and specify the file format. For this worked example, you can download the GBM data set from: http://reactomews.oicr.on.ca:8080/caBigR3WebApp/Cytoscape/GBM_genelist.txt.
5. Under FI Network Construction Parameters, select the “Fetch FI annotations” and “Use Linker genes” features, and then click OK (*see* **Note 19**). The constructed FI network based upon the uploaded gene list will be displayed in the Cytoscape network view panel (Fig. 4).
6. To query detailed information on selected FIs (*see* **Note 20** for details on the different types of FIs displayed in the network), select the FI edge of interest. For example, select the line connecting the TP53 and MMP13 nodes, right click to invoke the pop-up menu as the cursor is hovering over the line. Choose under the “Reactome FI” option, the “Query FI Source” feature to display the summary of the source interaction information. In the pop-up window, select the line under “Reactome Sources” to display more detailed interaction data (*see* **Note 21**).
7. To identify “topologically unlikely” clusters (or groups of genes that are closer to each other on the network than you would expect by chance), run the network clustering algorithm [21] on the displayed FI network. To do this, invoke the pop-up menu by right clicking an empty space in the network view panel, and under “Reactome FI” option, select the “Cluster FI Network” feature. Nodes in different network modules will be

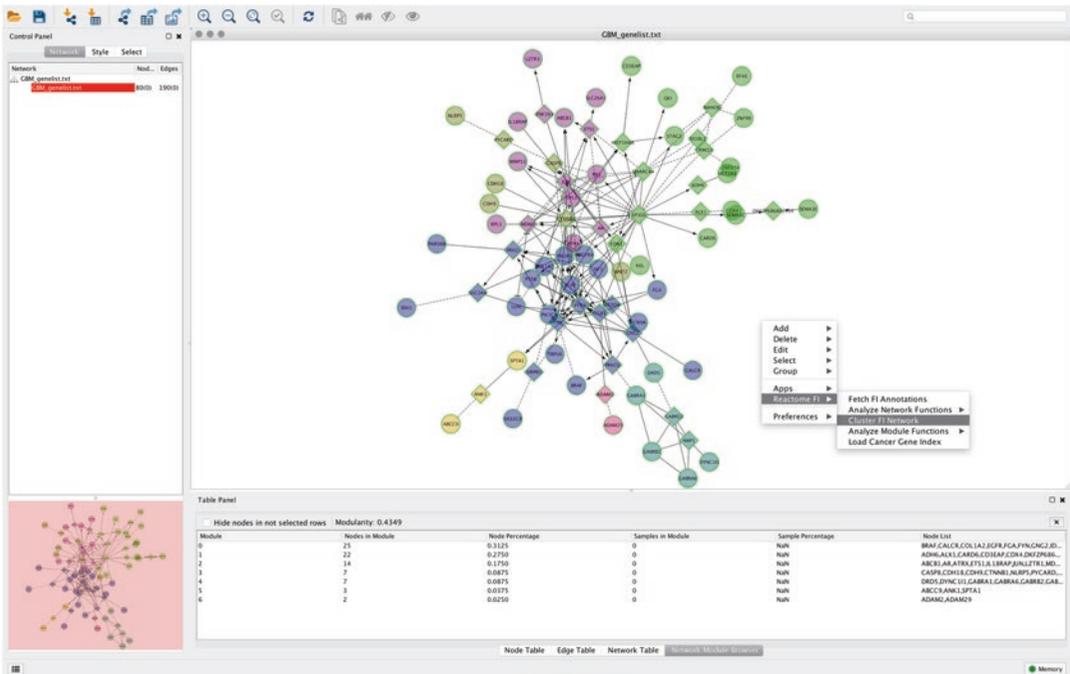


Fig. 4 The Reactome FI network created from the GBM gene list. An FI specific visual style will be created automatically for the FI network. The main features of ReactomeFIViz can be invoked from a pop-up menu, which can be displayed by right clicking an empty space in the network view panel

shown in different colors (*see Note 22*). Selecting a module from the table panel will yellow highlight the genes in the network view panel, which belong to the selected module.

8. To annotate the individual clusters with pathways, invoke the pop-up menu by right clicking an empty space in the network view panel, and under “Reactome FI” option, select the “Analyze Module Functions” feature, and click on “Pathway Enrichment.” To perform GO term enrichment analysis, select one of the GO term categories: GO Molecular Function, GO Biological Process, or GO Cellular Component from the “Analyze Module Functions” feature. Selecting a module size or FDR cutoff value will allow the user to filter enrichment results. After the analysis is complete, a new table panel will appear displaying the results of the pathway enrichment analysis (Fig. 5).
9. To overlay NCI Cancer Gene Index annotation onto the nodes within the network, invoke the pop-up menu by right clicking an empty space in the network view panel, and select “Load Cancer Gene Index.” In the Disease hierarchy panel, click through the different levels and terms, selecting and unfurling the relevant terms until reaching the lowest point of the hierarchy relevant to the query. Selecting a particular NCI Cancer

Module	GeneSet	RatioOfProteinGeneSet	NumberOfProteinGeneSet	ProteinFromModule	P-value	FDR	Nodes
0	Citoma(K)	0.0067	65	6	0.0000	<1.000e-03	EGRF, BRAF, PTEN, PDGFRA, PIK3CA, ...
0	Central carbon metabolism in ca.	0.0069	67	6	0.0000	<5.000e-04	EGRF, PTEN, PDGFRA, IDH1, PIK3CA, ...
0	Melanoma(K)	0.0073	71	6	0.0000	<3.333e-04	EGRF, BRAF, PTEN, PDGFRA, PIK3CA, ...
0	Prostate cancer(K)	0.0091	89	6	0.0000	<2.500e-04	EGRF, BRAF, PTEN, PDGFRA, PIK3CA, ...
0	Focal adhesion(K)	0.0212	207	7	0.0000	<2.000e-04	EGRF, BRAF, PTEN, PDGFRA, COL1A, ...
0	Endometrial cancer(K)	0.0053	52	5	0.0000	<1.667e-04	EGRF, BRAF, PTEN, PIK3CA, PIK3R1, ...
0	Signaling by FGFR(K)	0.0159	155	6	0.0000	<1.429e-04	EGRF, BRAF, PTEN, PDGFRA, PIK3CA, ...
0	PP3 activates AKT signaling(R)	0.0096	94	5	0.0000	<1.250e-04	EGRF, PTEN, PDGFRA, PIK3CA, PIK3R1, ...
0	Rap1 signaling pathway(K)	0.0216	211	6	0.0000	<1.111e-04	EGRF, PIK3CB, BRAF, PDGFRA, PIK3, ...
0	Signaling events mediated by PTP, ...	0.0053	52	4	0.0000	<1.000e-04	EGRF, PIK3CA, TRPV6, PIK3R1, ...
0	FoxO signaling pathway(K)	0.0136	133	5	0.0000	<9.091e-05	EGRF, BRAF, PTEN, PIK3CA, PIK3R1, ...
0	Signaling by SCF, KCTD(K)	0.0140	117	5	0.0000	<8.333e-05	EGRF, PTEN, PIK3CA, PIK3CA, PIK3R1, ...

Fig. 5 The table panel displaying the results of a pathway enrichment analysis

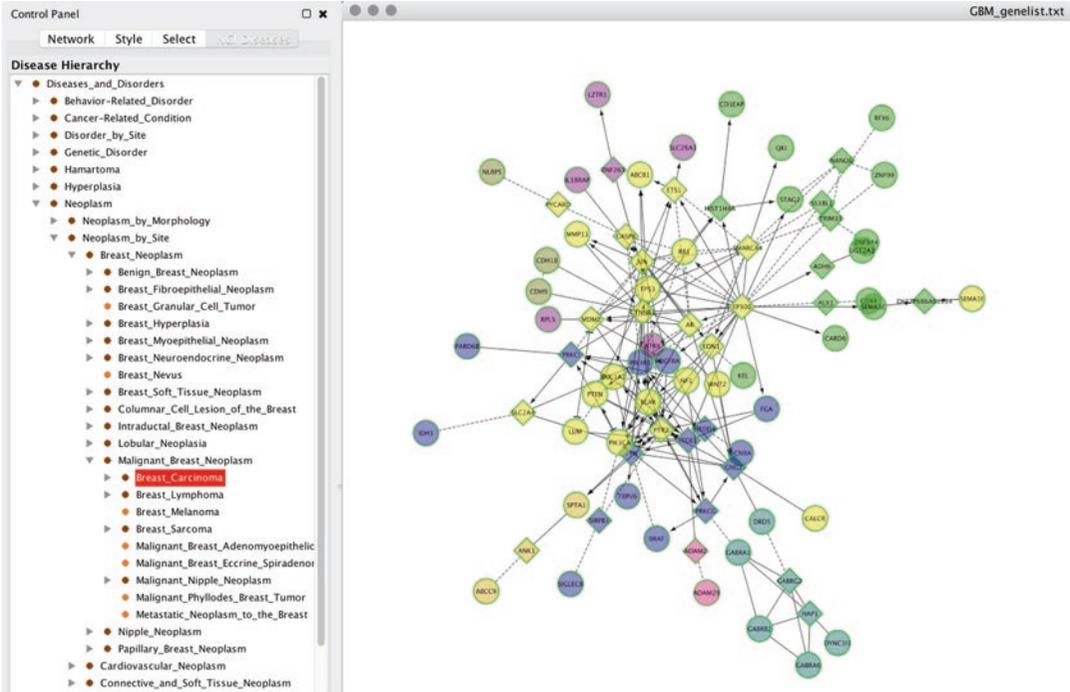


Fig. 6 The NCI Cancer Gene Index feature overlays yellow highlighting on genes in the network that have breast carcinoma annotations

Gene Index annotation term will yellow highlight the genes in the network, which have that particular annotation (Fig. 6).

4 Notes

1. *Values in the configuration file.* In Reactome, we refresh our FI network once per year. Therefore, the value for “YEAR” should be set as the current year (e.g., 2015 for the 2015 version). RESTUL_DIR is usually set as “results/{YEAR}”. The {YEAR} parameter will be replaced by the value of “YEAR” automatically by the program. The value of DATA_SET_DIR is usually “datasets.”

2. *Log4j.properties*. There are two rootLoggers configured in the log4j.properties. It is recommended that you choose LOGFILE so that the log can be output to a local file. However, you may use the console output (A1) for some methods to view output in the Eclipse console view. It is suggested that you copy all log output from Eclipse to a file for future reference. A file called Combined_Logging.txt is created for this purpose.
3. *Suggested names for data source files and output files*. You can use existing file names in configuration.prop as examples. Usually you can just change dates in the directory names for newly downloaded files.
4. *Loading mysql dump into mysql*. In order to make sure mysql can find the dump file, start mysql from the directory containing the unzipped mysql dump file. Otherwise, you have to use the absolute path to your dump file.
5. *Check updated database schema*. To check if your database schema has been updated, download a copy of the Reactome curator tool from <http://www.reactome.org/download>, connect the tool to your database using menu Database/Database Browser/Schema View, and check if you can find two classes, Interaction and TargetedInteraction under the Event class, and an attribute called dataSource for the topmost class, DatabaseObject. To get the latest version of the database schema for the Reactome API used in multiple places, choose menu File/Export Schema to overwrite the original copy of schema in the resources folder.
6. *NCI-PID database*. The NCI-PID pathway was retired around the end of 2015. Its content has been hosted by the NDEX database. We have not tried using the NDEX database, and believe you may skip pathways from this source to construct a functional interaction network with enough coverage since the coverage in Reactome is much higher at present and can cover the loss from the NCI-PID pathways.
7. *ENCODE TF/target interactions*. There are many TF/Target interactions in the original ENCODE project release. These interactions were detected based on ChIP-seq (chromatin immunoprecipitation sequencing) and many of them may not play actually biological roles inside cells. We use a simple filter to choose TF/Target interactions that are supported by gene co-expression and/or GO BP annotation sharing.
8. *Loaded Ensembl-Compara database*. To make sure the procedures used to load the content for the database worked as expected, use the following query after logging into mysql and selecting your database:

```
SELECT COUNT(*) FROM seq_member WHERE taxon_id = 559,292 AND source_name LIKE 'UniProt%';
```

The returned value should be around 6200.

9. *Gene co-expression data files.* The original downloaded data files contain gene pairs only. In order to be used as features in NBC, we have mapped gene names to UniProt accession numbers. We used the downloaded UniProt data file to do this mapping. We used the SwissProt part of the UniProt data for doing the mapping for the Prieto data file. For Lee's data file, we used the original downloaded mapping file, refseq-hs-annts.txt, chose correlations generated from three or more data sets, and then normalized with the latest UniProt data.
10. *Method convertPathwayDBs.* In order to keep the log file, make sure log4j.prop is configured to write output to a file and assign enough memory for running this method (e.g., -Xmx4G). Some of identifiers used by several databases cannot be mapped to UniProt accession numbers, including a very small number (less than 200) of KEGG ids from KGML pathways, as well as some of gene names used by TRED and ENCODE. You may see a FileNotFoundException at the end of the method run. You can just ignore it.
11. *Method dumpPathwayDBs.* It is recommended that you connect your curator tool to the database and check instances converted from other non-Reactome sources by searching based on the dataSource attribute. Some UniProt accession numbers may be duplicated if they cannot be mapped to Reactome ReferenceGeneProduct instances and are used in multiple data sources. This should be fine for generating the FI network after normalization.
12. *Method trainNBC.* After finishing running this method, you should also run class NBCGUITest to study the prediction results using different combinations of features (*see* Fig. 7 for a screenshot from running this class).
13. *Time to construct the Reactome FI Network.* All results should be output into the folder defined by property RESULT_DIR in the configuration file. The whole process including collecting all data sources usually takes 2 days with most of the time spent on collecting data files. The time used for Section 3.5 should be around half a day.
14. *Cytoscape installation.* You can install Cytoscape from a compressed archive distribution, from source, or use one of the automated installation packages that exist for Windows, Mac OS X, and Linux platforms.
15. *Launching Cytoscape.* The instructions for this will depend on your computer's operating system. For Mac or Linux, double click on the Cytoscape icon in the Application/Installation folder. For Windows, open the Cytoscape folder through Start Menu, and then click on the Cytoscape icon. The Cytoscape desktop and the welcome screen should now appear.

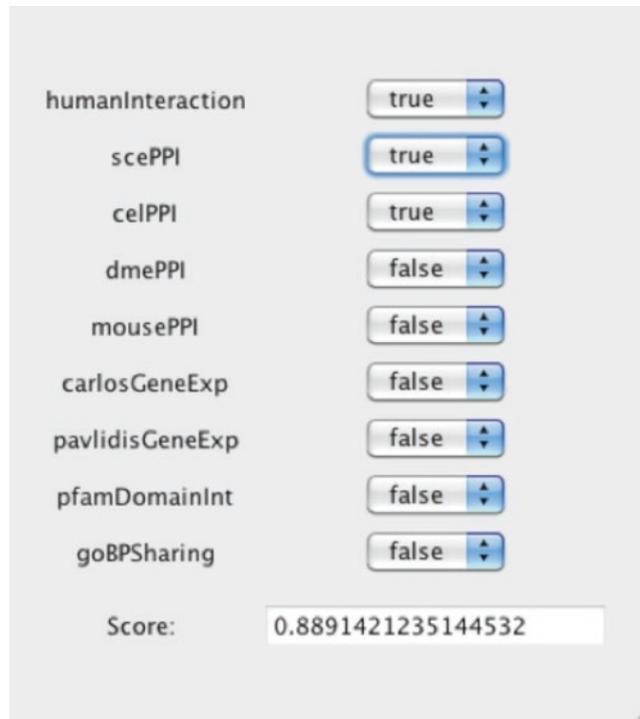


Fig. 7 Screenshot of the window generated by running class NBCGUITest

16. *Alternative ReactomeFIViz App Installation.* The reader can also directly install the ReactomeFIViz App from the Cytoscape App Store web site:
 - (a) Launch Cytoscape software
 - (b) Use a browser to load the Reactome web within the Cytoscape App Store page: <http://apps.cytoscape.org/apps/reactomefiplugin>
 - (c) Click on the “Install” button. A dialog window should pop out showing the progress.
 - (d) When installed, the button on the ReactomeFI plugin web page will change to “Installed.”
17. *ReactomeFIViz supported file formats.* Three different file formats are supported for gene set/mutation analysis: (1) Simple gene set with one line per gene; (2) Gene/sample number pair, which contains two required columns—gene and number of samples having the gene mutated, and an optional third column listing sample names (delimited by “;”); and (3) NCI MAF, which is the mutation file format used by The Cancer Genome Atlas project (<http://cancergenome.nih.gov/>). The HotNet Mutation Analysis, for doing cancer mutation data analysis, uses the NCI MAF format. ReactomeFIViz can also load a pre-normalized gene expression data file. In this case, an expression data file should be a tab-delimited text file with table

headers. The first column should be gene names. All other columns should be expression values in different samples.

18. *Reactome FI Network Version*. It is possible to get different results using different FI network versions because a later version may contain more proteins/genes and more FIs. From our experience, however, a significant FI network module is usually stable across multiple versions.
19. *FI Network Construction Parameters*. The “Fetch FI annotations” feature will display the FI attributes on the network. “Use Linker genes” will add “linker” genes to the network, which are not part of the uploaded gene list but known to interact with members of the gene list, increasing the connectivity within the network.
20. *FI Edge Attributes*. Edges will be displayed based on FI direction attribute values. In Fig. 4, “->” for activating/catalyzing, “-|” for inhibition, “-” for FIs extracted from complexes or inputs, and “---” for predicted FIs. See the Cytoscape “VizMapper” tab, “Edge Source Arrow” Shape, and “Edge Target Arrow” Shape values for details.
21. *MMP13-TP53 FI Source Information*. For the worked example, the MMP13-TP53 interaction is derived from TRED, where TP53 is the transcription factor with MMP13 the target, and is supported by functional analysis (i.e., Western and Northern blots) and literature citations (i.e., PMID 10753945, PMID 10415795).
22. *Clustering Results*. In the worked example, there are 7 modules containing 2 genes or more, with the largest modules containing 25 genes. Different colors are used only for first 15 modules based on their sizes.

References

1. Rual J-F, Venkatesan K, Hao T et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437:1173–1178
2. Ewing RM, Chu P, Elisma F et al (2007) Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol Syst Biol* 3:89
3. Fabregat A, Sidiropoulos K, Garapati P et al (2016) The Reactome pathway knowledge-base. *Nucleic Acids Res* 44:D481–D487
4. Gerstein MB, Kundaje A, Hariharan M et al (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489:91–100
5. Jiang C, Xuan Z, Zhao F et al (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* 35:D137–D140
6. Wu G, Dawson E, Duong A et al (2014) ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *FI000Res* 3:146
7. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504
8. Wu G, Feng X, Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* 11:R53
9. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43:D204–D212
10. McGarvey PB, Huang H, Barker WC et al (2000) PIR: a new resource for bioinformatics. *Bioinformatics* 16:290–291

11. Kanehisa M, Sato Y, Kawashima M et al (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44:D457–D462
12. Schaefer CF, Anthony K, Krupa S et al (2009) PID: the pathway interaction database. *Nucleic Acids Res* 37:D674–D679
13. Mi H, Poudel S, Muruganujan A et al (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 44:D336–D342
14. Razick S, Magklaras G, Donaldson IM (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinf* 9:405
15. Lee HK, Hsu AK, Sajdak J et al (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res* 14:1085–1094
16. Prieto C, Risueno A, Fontanillo C et al (2008) Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS One* 3:e3911
17. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25:25–29
18. Flicek P, Aken BL, Ballester B et al (2010) Ensembl's 10th year. *Nucleic Acids Res* 38(Database):D557–D562
19. Finn RD, Coggill P, Eberhardt RY et al (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44:D279–D285
20. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455:1061–1068
21. Newman MEJ (2006) Modularity and community structure in networks. *Proc Natl Acad Sci U S A* 103:8577–8582

Prediction of Protein Interactions by Structural Matching: Prediction of PPI Networks and the Effects of Mutations on PPIs that Combines Sequence and Structural Information

Nurcan Tuncbag, Ozlem Keskin, Ruth Nussinov, and Attila Gursoy

Abstract

Structural details of protein interactions are invaluable to the understanding of cellular processes. However, the identification of interactions at atomic resolution is a continuing challenge in the systems biology era. Although the number of structurally resolved complexes in the Protein Databank increases exponentially, the complexes only cover a small portion of the known structural interactome. In this chapter, we review the PRISM system that is a protein–protein interaction (PPI) prediction tool—its rationale, principles, and applications. We further discuss its extensions to discover the effect of single residue mutations, to model large protein assemblies, to improve its performance by exploiting conformational protein ensembles, and to reconstruct large PPI networks or pathway maps.

Key words Structural matching, PPI prediction, Mutation mapping, PPI network, Structural pathway modeling

1 Introduction

1.1 Review of Template-Based Approaches

In the cell biological processes are realized through interactions among proteins and between proteins and other molecules. Protein–protein interactions (PPIs) orchestrate complex processes including signaling and catalysis. With the recent advances in experimental techniques, the number of identified PPIs keeps increasing. However, we are still far from complete knowledge of PPIs and their characterization at the atomistic level. Template-based computational methods to predict PPIs have recently become popular due to the significant increase in protein sequence and structural data. In this chapter, we review template-based approaches, not only to predict PPIs but also to provide structural models of the interacting proteins. Here, we describe PRISM, one of the earliest template-based algorithms that aimed to build PPI networks (called

structural PPI networks) based on augmented sets of structural models of the interactions. The integration of protein structures into the PPI network allows mapping of single nucleotide variations (SNVs) in protein-coding areas to interactions and predicting the effect of the mutations in larger settings. The fundamental concept behind PRISM is that there are favorable structural motifs at protein–protein interfaces and that these architectural motifs resemble those found in protein cores [1–5].

1.2 *Template-Based Docking*

Template-based computational approaches for predicting PPIs and their structural models utilize the accumulated sequence and structure data of known PPIs [2]. Given two unbound proteins, the task of the template-based method is to find a similar complex (template) in the database of known interactions, aligning the unbound proteins to the template. One of the earliest approaches is the homology-based prediction of PPIs [6], which is based on the observation that homologous pairs of proteins tend to interact in a similar manner. The homology-based method requires significant sequence similarity (at least 30 %) and uses the whole sequence. Interactome3D is among the homology-based methods. Additionally, the domain–domain preferences of protein interactions are considered as partial templates in Interactome3D [7]. The Instruct method also uses domain-mediated interaction templates and searches for domain availability in target proteins [8]. In the case of low sequence similarity, a protein threading approach can be used [9, 10]. This method was initially used for predicting the structure of single proteins, and later the methodology has been extended to model PPIs [9, 10]. In both cases, similarity between the unbound proteins and templates as a whole is sought. This limits the applicability at large scale due to many factors such as low coverage of available data, conformational changes of the proteins, and so on.

Another template-based approach is to use partial structures as templates, where the most representative case is the use of protein interfaces. PRISM is one of the first methods that uses protein interfaces to predict PPIs and their structural models [11, 12]. In this chapter, we outline the basics of PRISM and its extensions for PPI prediction. We include *in silico* analysis of functional variations, constructing large assemblies (*see Note 1*) and modeling PPI networks in atomic detail. We also detail the functionalities of the PRISM web server for online prediction.

2 Methods

2.1 *Principles of PPI Prediction by PRISM*

PRISM combines sequence and structural information about known protein interfaces to discover not only potential novel interactions but also the binding modes of known protein interactions. The method is well established, accurate, and computationally efficient. PRISM serves both as an online resource [13] and a

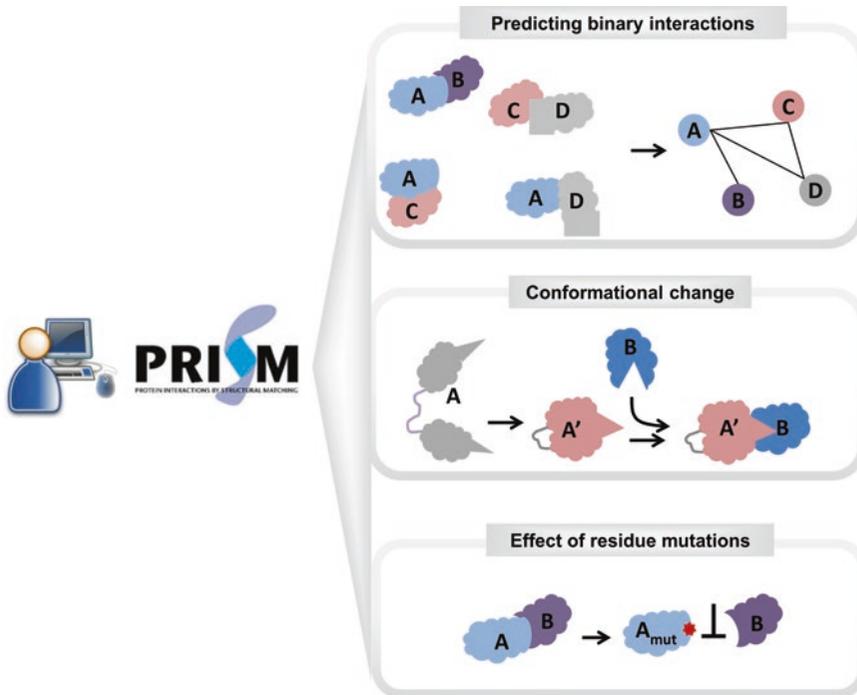


Fig. 1 A schematic representation of PRISM functionalities. PRISM can be run on local computers as well as interactively on the web server. PRISM predictions can be used for multiple purposes. Predicting the interaction between two proteins and PPI networks is possible with PRISM. In addition, with an accurate design of the template set, conformational changes can be handled in the prediction. The effect of residue mutations can be analyzed by comparing the wild-type and mutated predictions

downloadable protocol [12] (Fig. 1). To run the PRISM system, two datasets are necessary: (1) a template set and (2) a target set. The idea is that if partner chains of a template interface are spatially similar to any region on the surfaces of the two targets and share some evolutionarily important residues, then these target proteins can interact with each other with an architecture resembling that of the template interface.

2.1.1 Template Set

The template set is composed of known protein interfaces. Protein complexes deposited in the Protein Data Bank (PDB) are the main resource to extract protein interfaces. The performance of PRISM depends on the diversity of the interfaces in the template dataset. The ideal set should cover all available architectures of protein interfaces. An approximate template set can be formed by clustering structurally similar interfaces and choosing one member from each cluster. Details on how to construct an interface dataset using this approach is discussed in [14–16]. PRISM provides a built-in (default) template set prepared in this way, which is a subset of the protein interfaces in the PDB. Depending on the system under investigation, the template set can be modified—e.g., using only

oncogenic interfaces to model cancer-specific interactions. The latest version of the interface dataset used in PRISM was constructed from the all PDB complexes deposited before January 2012. This version includes 22,604 unique interface architectures.

In addition to structural similarity, PRISM uses matching of some critical residues. The residues in protein interfaces do not equally contribute to the binding energy of interactions. Some interface residues are more important; these are called “hot spots”. Experimentally, they can be determined by alanine-scanning mutagenesis experiments, but these data are available only for a limited number of complexes. Therefore, computational approaches emerged to accurately and efficiently predict binding hot spots. Hotpoint, which considers solvent accessibility and contact potentials of residues [17, 18], is one of these approaches. If a residue is highly packed and buried in the interface, it has a higher tendency to be a hot spot. Hot spots predicted by the Hotpoint server are used in the built-in template set of PRISM. However, the user is free to label hot spots with other available methods.

2.1.2 Target Set

The target set is composed of all proteins under consideration. It should contain at least two proteins. The atomic coordinates of the targets are retrieved from the PDB (for details *see Note 2*). For proteins not available in the PDB, homology modeling-based techniques can be used to increase the structural coverage of the target set.

The target set can be shaped based on the purpose of the analysis. If the aim is to discover a specific interaction between two proteins, the target set should contain all available structures of the two proteins. To structurally model a known pathway, all proteins functioning in that pathway form the target set. Another possibility is to construct an organism-specific structural interactome with PRISM. In this case, all proteins having structural information or homology models in that organism are included in the target set. If the aim is to construct a tissue-specific structural interactome, the target set should cover all proteins expressed in the selected tissue.

Recent efforts using PRISM have shown that considering all available 3D conformations of proteins improves the accuracy of prediction [19] (*see Note 3*). Proteins are flexible and prone to conformational changes. Binding of an activator to the extracellular portion of a membrane protein or binding a small molecule to a region of a protein that is distant from the functional region can lead to allosteric changes in global structure. Domain motions including hinge motions also reflect conformational changes, typically on a larger scale. For example, when a small molecule (trifluoperazine) binds to calmodulin, calmodulin adopts a new conformation that opens up two helices and changes its binding preferences. Even

binding of an ion (Ca^{2+}) shifts the equilibrium between the open and closed calmodulin states. The PDB is a rich source for multiple protein conformations.

2.1.3 Prediction

The prediction procedure follows four consecutive steps: (1) extraction of the surface regions of target proteins, (2) structural alignment of the templates to the targets, (3) transformation of the targets onto the templates and filtering unrealistic cases, and (4) flexible refinement and energy calculation.

Proteins interact using their surfaces. Therefore, the first step is extracting the atomic coordinates of target protein surface residues. The NACCESS [20] tool, which calculates the solvent accessibility of protein residues, is used for this purpose. Residues are defined as being on the surface if the ratio of their solvent accessible surface area in the protein state and in an extended tripeptide state is greater than 15 %. To conserve the secondary structure, residues within 6 Å of the surface are also extracted. These are called “nearby” residues. The output of this step is the atomic coordinates of the surface and nearby residues of each of the target proteins in PDB format.

The next stage is rigid body structural alignment, which is sequence order independent where only geometric similarity is sought. PRISM uses Multiprot [21] for structural alignment. At the alignment stage, conservation of the template interface hot spots on the target surface is also checked to limit the search space. If the surfaces of two targets have spatially similar regions to complementary partners of a template interface and at least one hot spot in each partner chain is conserved on the target surfaces, the global structures of the targets are superimposed onto the template. In this way, the first putative complex is modeled. The superimposition of the targets may cause atomic clashes. At the filtering stage, predicted complexes having many atomic clashes are removed from the final list. The last stage entails refinement of predicted complexes and ranking them according to their binding affinities. Fiberdock [22] is employed for refinement and energy calculation purposes. First, the side chains in the interaction interface and the backbone of the predicted complex are optimized. The binding energy score is also calculated. In this way, PRISM considers not only geometric complementarity but also the flexibility of targets and their chemical complementarity. The final list gives a set of predicted complexes refined with docking protocols.

After completion of all intermediate stages, the final result includes the list of putative binary interactions, their binding energy scores, and their 3D modeled complexes. The output is a rich resource for further analysis and modeling. PRISM is multi-functional. It can be used for predicting binary interactions, discovering the effect of residue mutations, constructing protein

interaction networks, and structural modeling of known pathways (Fig. 1). Each of these functionalities is reviewed in the following sections.

2.2 PRISM2.0 Web Server

PRISM can be used in two ways: (1) PRISM stand-alone version [12] or (2) PRISM web server [13]. Advanced users may prefer to install the PRISM protocol and use it on their local computers. This allows adjustment of parameters for prediction or getting output for intermediate steps. The stand-alone version runs only in the Linux environment. PRISM is continuously upgraded to improve its performance and user-friendliness as well as its computational effectiveness. The current stand-alone protocol (version 1.0) can be accessed at http://prism.ccbb.ku.edu.tr/prism_protocol/. An upgraded stand-alone version will be released soon. The PRISM web server, on the other hand, has been developed to provide a simple user interface to perform all the steps of the PRISM algorithm with default settings. The web server (PRISM2.0) [13] implements a slightly modified version of the stand-alone protocol [12]. In addition, the web server aims to provide a repository (database) of structural models of all predicted interactions. All the predicted interactions between proteins having PDB IDs are stored in its database. As the community uses the server, it is expected that the structural models in the database will grow and become an invaluable resource.

The web server is multifunctional, which allows for online prediction as well as browsing the accumulated data. We provide an overview of the web server in Fig. 2. The step-by-step procedure to run the prediction algorithm on the PRISM server is provided in Subheading 2.2.1.

2.2.1 Online Prediction

The user first needs to prepare the target pairs and determine the templates to be used in the prediction.

1. Because PRISM requires the structures of target proteins, protein names or any other identifiers must be cross-referenced to the PDB identifiers. As mentioned in **Note 2**, one protein may have many PDB identifiers; each may represent partial structures of different fragments or structures at different resolutions. For example, human RAF protein has many structures in the PDB (e.g., chain B of 1c1y covers positions 55–131, chain A of 1faq covers positions 136–187, and chain A of 3omv covers 323–618). All structures of the corresponding target protein should be included for a better accuracy. As input, the user can provide the PDB codes of these proteins with or without chain identifiers or upload the PDB formatted files of the target structures. If chain identifiers are included in the PDB id, then only those chains will be used. For example, the input 1a0hAE will use only the A and E chains (skipping the other chain in 1a0h) to form the protein structure.

A



B

Predict Interactions

Two Proteins Network

Two Proteins:

Target1: or Upload Target1 No file chosen

Target2: or Upload Target2 No file chosen

Template(optional):

Email(optional):

Predict Interactions

Two Proteins Network

Network:

Paste pair-list: or Upload pair list No file chosen

Template(optional):

Email(optional):

C

Target1	Target2	Interface	Energy	Structure
2j5x	3a6pC	2j5xAB	-12.98	<input type="button" value="View"/>
2j5x	3a6pC	2j5xAB	-0.63	<input type="button" value="View"/>

View Structure



Contacts of Interface Residues Structure

Interface Residues	Contacts	target2_chain_resName_resNo
target1_chain_resName_resNo	<->	target2_chain_resName_resNo
2j5x_B_ASP_92	<->	3a6pC_C_PRO_49
2j5x_B_VAL_39	<->	3a6pC_C_LEU_43
2j5x_B_ASP_94	<->	3a6pC_C_VAL_51

Fig. 2 Overview of the PRISM web server. (a) The “Prism” tab is designed for online prediction. The “Predictions” tab is for browsing the accumulated data in the web server. The “Templates” tab is for browsing detailed information about the built-in template set. (b) For online prediction, two options are available: (1) predicting the interaction between two proteins (*left panel*) or (2) predicting interactions in a network (*right panel*). For the former option, the input is the PDB codes or PDB files of “Target 1” and “Target 2.” For the latter option, the input is the list of protein pairs with their PDB codes. The template code and e-mail address are optional inputs for both. (c) The “Results” page lists the predicted interactions with the target codes, the template (the column labeled “Interface”), and the calculated binding energy. Each interaction has a “View” button where an interactive visualization of the predicted complex is possible. Target 1 is colored *blue* and Target 2 is *red*. Their interface regions are in *pink* and *light blue*, respectively. The contacts in the interface region and the PDB file of the predicted complex are downloadable using the “Contact of Interface Residues” and “Structure” links, respectively

2. If the protein does not have information in PDB, homology modeling techniques can be used to predict the structure. Some useful resources for homology modeling are MODELLER [23], SWISS-MODEL [24], and ModBase [25]. The input for online prediction in PRISM would then be the PDB formatted files of the homology models.
3. With its default settings, PRISM uses the built-in template set. However, the user can enter a template name to run PRISM only for one interface. This option is useful if the user is interested in predicting an interaction using a specific template (much faster than using the whole template set) or in using a template interface that is not available in the built-in set.
4. There are two prediction options (Fig. 2b): the first is to search for the interaction between two proteins and the second is to predict a network of interactions for up to 10 pairs of proteins. If the target set contains a structure prepared by the user with homology modeling, then only the “predict between two proteins” option can be used.
5. After preparing the target pairs and determining the templates to be used, the job can be submitted. The user can wait on the page, bookmark the page to browse later, or enter an email address to receive a notification after the job is completed.
6. For each prediction, the PDB codes and chain identifiers of the targets, the template interface used in the prediction and the calculated binding energy are listed (Fig. 2c). Also, a “View” button is available for interactive visualization of the predicted complex in Jmol where user can examine the predicted interface as well as noninterface regions in the overall complex. The PDB formatted file of the predicted complex and contacting residues in its interface are downloadable on the results page.

2.2.2 Browsing Accumulated PPI Predictions in PRISM

The PRISM web server also provides a search option and a download option to retrieve predictions from the accumulated data.

1. On the upper panel (shown in Fig. 2a), the user can select the “Predictions” tab to browse the PPIs already predicted and deposited at PRISM. Only the predictions for structures available in PDB are stored in the database. User-supplied protein structures (i.e., from homology modeling) are not accumulated. All the functionalities including visualization and download are also available for the accumulated data.
2. In addition, the user can search the data either with the target code or with the template interface code. For example, the target 3g9wD has five predicted interactions in the accumulated data and these interactions are predicted from the templates 3g9wAD and 1jfiBC. When we search the data with a template interface code, e.g., 3fxdBD, we find that there are 38 predicted interactions based on this interface.

2.2.3 Analysis of Template Interfaces

The built-in template dataset can be also browsed and analyzed in detail.

1. On the upper panel (shown in Fig. 2a), the user can select the “Templates” tab to browse the built-in template set. PRISM provides a table listing the representative of each cluster of interfaces (“Representative” column) as well as the number of interfaces in the cluster (“Members” column).
2. Clicking on an interface in the “Representative” column displays a page in the HotRegion database where the user can find the computational hot spots and hot regions.
3. Clicking on a number in the “Members” column opens a new window listing the interfaces structurally similar to the representative interface. Each member interface has a web link to the HotRegion database.

The PRISM web server has also a “Tutorial” page where the definitions of all terms are mentioned and the principles of prediction procedure can be found in detail (Fig. 2a).

2.3 Mapping Residue Level Mutations onto Predicted Binding Sites

The reconstructed structural PPI networks of signaling pathways help complete missing parts of pathways and provide PPI details. For example, they reveal the details of how signals coming from different upstream pathways merge and propagate downstream, how parallel pathways compensate for each other, and how multi-subunit signaling complexes form. These networks can also offer the possibility of observing the effects of SNPs and mutations on signal propagation.

There has been a tremendous increase in genomic variations data as well as in PPI data. Genome-wide association studies, whole-genome sequencing, and exome sequencing have shown that each personal genome contains millions of variants, thousands of which are nonsynonymous single nucleotide variants (nsSNVs) causing changes at the residue level. Some of these variations are neutral, whereas some are disease associated. For example, cancer results from somatically acquired variations in the DNA of cancer cells. However, not all the somatic changes present in a cancer genome are involved in development of the cancer. Indeed, it is likely that some have no contribution at all. These are termed as passenger mutations, whereas the ones that do contribute to disease are driver mutations. The identification and characterization of disease-associated, driver variations are important tasks in personalized medicine and genomic sequencing [26]. Genetic variation and mutation data are publicly available in databases such as the Online Database of Mendelian Inheritance in Man (OMIM) [27], the Human Gene Mutation Database (HGMD) [28], Humsavar (<http://www.uniprot.org/docs/humsavar>) of Uniprot, ClinVar [29], COSMIC (Catalogue of Somatic Mutations in Cancer) database [30], and cBioPortal for Cancer Genomics (which contains

data from the Cancer Genome Atlas (TCGA) project) [31]. These databases capture variants at the level of the gene nucleotide sequence as well as the corresponding amino acid changes in the protein (gene product). These variations can be polymorphisms, variations between strains, isolates, or cultivars, disease-associated mutations, or RNA editing events. Most databases provide detailed information about the mutations, including the mutated residue numbers.

In our previous studies [32, 33], we extracted available point mutations (and data related to these mutations, e.g., phenotypic effects) from different sources and modeled the effects of oncogenic mutations. First, we identified mutations that mapped to the interface region of our modeled protein–protein complexes. Next, we performed *in silico* mutagenesis to observe the contributions of these specific mutations to the interaction (see Subheading 2.5). We re-ran PRISM on the mutant structures [32, 33] and modeled the new interaction between the mutant target and its partner. PRISM results for the wild-type and mutated cases give insight into the edgetic effects of the analyzed mutation. A mutation may cause a loss of an interaction, cause a gain, or be neutral. In Fig. 3, we show a conceptual representation of edgetic effects in a PRISM predicted network. A hypothetical protein “A” having two binding sites (*red* and *blue* colored regions) loses three interactions when a mutation occurs in one of its binding

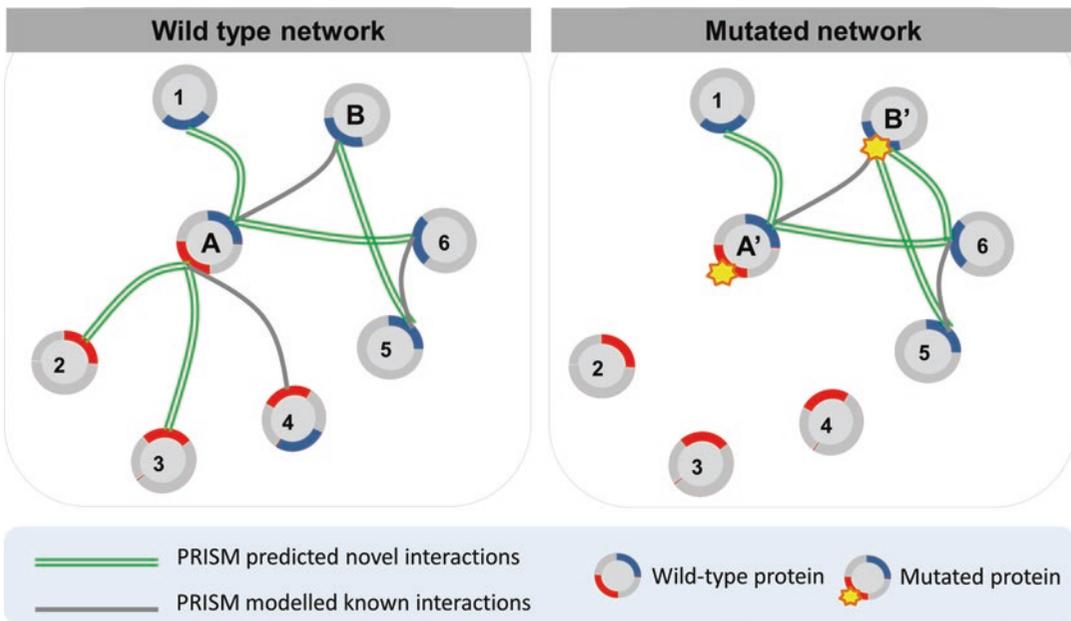


Fig. 3 A representation of the concept of a mutation effect in a PRISM predicted network. Protein A has two binding sites (*red* and *blue* colored regions) and the mutation occurs in the red binding site in A' (mutated protein A). Known interactions are *gray* and novel interactions are *green double lines*. A mutation in protein A leads to loss of interactions between A and P2, P3, P4. On the other hand, when the mutation occurs in protein B, it gains a new interaction (with P6) while conserving its wild-type interactions

sites. However, a mutation in the binding site of hypothetical protein “B” is neutral for its wild-type interactions. However, a new interaction is gained when this mutation occurs in B. This type of analysis has been applied to the interleukin 1 (IL-1)-initiated signaling pathway [32]. Mapping residue mutations onto the reconstructed structural IL-1-initiated signaling pathway improved our understanding of the activation/inhibition mechanism.

2.4 From PRISM Predictions to Interaction Networks and Pathway Modeling

PRISM predicts PPIs at the proteome scale in a computationally efficient way, which enables us to construct large PPI networks. In these networks, nodes represent proteins and edges represent the interaction between them. The output of PRISM is the PDB codes of the chains of predicted interacting pairs, their binding energy, the structural model of the predicted complex, and the name of the template interface. To convert the PRISM output to a classical PPI network, each structural state is mapped to its corresponding protein name (*see Note 4*). The Uniprot database can be used as the central cross-referencing resource to map PDB chains to their unique protein identifiers. To illustrate a sample network predicted by PRISM, we compiled all the interactions accumulated in the web server. Since 2014, 542 unique human protein targets from 1193 PDB chains have been tested by external users. In total, 749 unique interactions have been predicted (Fig. 4). Among them, 246 are known interactions supported by experimental or database evidence in the STRING database [34] with a confidence score

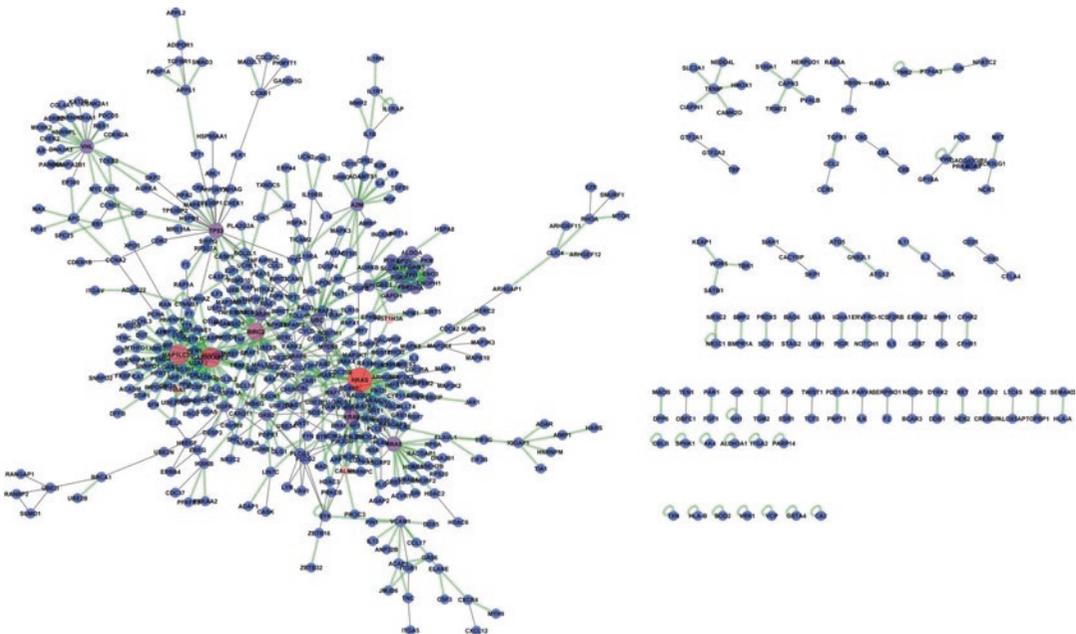


Fig. 4 A network constructed with the interactions that have been predicted by users of the PRISM web server. Node sizes are scaled based on their centrality in the network. *Gray edges* are PRISM predicted and STRING validated interactions. *Green two-line edges* are PRISM predicted novel interactions

greater than 0.80. The remaining 503 interactions are novel—discovered by PRISM. The sizes of the nodes are scaled according to their centrality in the network. The node colors change from blue to red based on their centrality.

Additionally, a known pathway map can be structurally reconstructed by PRISM predictions. Adding the structures of protein interactions is necessary for rational pathway modeling and improves the understanding of how a function is exerted or a signal is transduced in a pathway. In previous work, PRISM has been applied to structurally model several pathways. Among them, the reconstructed p53-mediated pathway has been used to illustrate the multipartner interaction preferences of hub proteins including p53 and Mdm2 [35].

In general, pathway maps are incomplete. Predicted novel interactions help in reconstructing the pathway and in elucidating unknown functionalities. The human ubiquitination pathway, IL-1-initiated signaling pathway, MAPK pathway, and Toll-like receptor signaling pathway have also been structurally modeled with PRISM. For this type of analysis, the initial target set is all PDB chains of the proteins present in the pathway of interest. The output is structural models of known interactions and predicted novel interactions. By superimposing all partners of a protein, overlapping and distinct binding sites can be identified. This enables discriminating simultaneous or mutually exclusive interactions.

2.5 Case Study

PRISM can also be utilized to reconstruct phenotype-specific subnetworks of PPIs. Previously, we formed subnetworks using the critical seed genes for two different phenotypes: lung and brain metastasis from primary breast tumors [33]. The seed genes were obtained from the literature [36, 37]. We extended the networks around these genes by using GUILD [38]. Finally, we modeled the structures of protein–protein complexes in these subnetworks and mapped the mutations from COSMIC [30] to the protein interfaces. This permitted the evaluation of the effect of mutations on the interactions. For example, EGFR was observed to make different interactions in the two reconstructed subnetworks. In brain metastasis, EGFR was found to interact with HBEGF, whereas the same protein was found to interact with EREG in lung metastasis. Indeed, HBEGF and EREG were known to have roles in brain [36] and lung [39] metastasis of breast cancer, respectively. When we examined the modeled structures of EGFR–HBEGF and EGFR–EREG complexes (Fig. 5), we realized that both complexes' interfaces included different genetic variations obtained from the COSMIC database [30]. The residue R377S mutation affects both EGFR–EREG and EGFR–HBEGF interactions. However, the D46N, Q432H, and V441I mutations were found only in the EGFR–EREG interface (*see Note 5* about how to determine residue positions from sequence to structure). Consequently, one can speculate that the latter three mutations

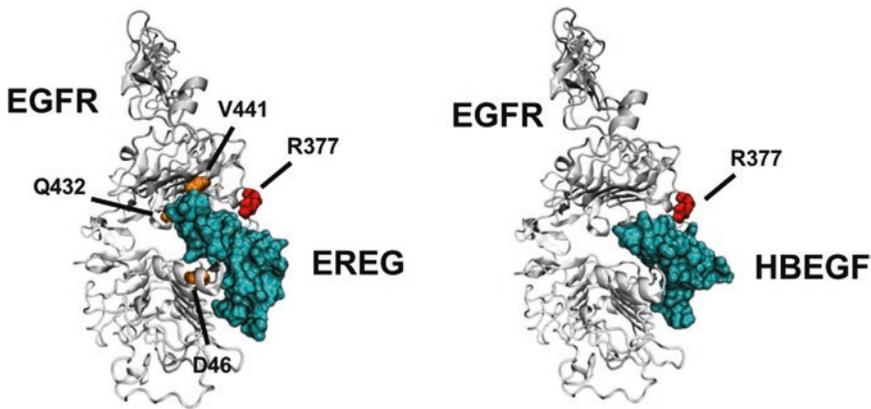


Fig. 5 The structures of EGFR-EREg and EGFR-HBEGF complexes modeled by PRISM

might have important roles in lung metastases initiating from breast tumors. These mutations may contribute in different ways to the stability of the complexes and thus to their signaling pathways. Some mutations can make the complex active all the time (a gain-of-function mutation), whereas others can make the interaction weaker (a loss-of-function mutation). Calculation of the effect of these mutations on the stability of the complex is needed to determine whether a mutation is likely to cause a loss- or gain-of-function.

3 Notes

The following are some points to be kept in mind:

1. Proteins often form large assemblies. Besides pairwise prediction of PPIs, discovering the organization of proteins in large assemblies is crucial to accurately model functional pathways. PRISM can be used to build three-dimensional models of protein assemblies from predicted binary interactions. PRISM has also been successfully applied to model symmetric cyclic structures and asymmetric complexes. Electron microscopy (EM) maps have been used for the validation of constructed assemblies. In a recent study, the TIR domain signalosome in the Toll-like receptor-4 signaling pathway has been constructed with PRISM predictions [40].
2. If the complete structure of the targeted protein is not available, all partial structures can be added to the target set (partial crystallization).
3. Using multiple conformations from the PDB increases the accuracy. Rather than using a single structure, structures captured under different conditions help to increase the conformational space.

4. A gene product might have several PDB codes. When preparing the target set and/or constructing PPI networks, mapping from gene codes to the corresponding protein codes needs to be done (possibly using UniProt).
5. The positions of mutations in the protein sequence may not match the residue position in the protein structure. To solve this problem, the fasta sequence of the protein structure can be aligned with the complete protein sequence and the residue positions can be relabeled. Another problem is that if a structure includes only a fragment of a protein, a particular position might not be present in the structure at all.

Acknowledgments

N.T. thanks to the TUBITAK-Marie Curie Co-funded Brain Circulation Scheme (114C026) and the Young Scientist Award Program of the Science Academy (Turkey) for the support. O.K and A.G. are members of the Science Academy (Turkey). We acknowledge the partial funding from TUBITAK projects (114M196 and 113E164). This project has been funded in whole or in part with Federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health, under contract HHSN261200800001E. This research was supported [in part] by the Intramural Research Program of NIH, Frederick National Lab, Center for Cancer Research.

References

1. Keskin O, Gursoy A, Ma B, Nussinov R (2008) Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chem Rev* 108(4):1225–1244
2. Muratcioglu S, Guven-Maiorov E, Keskin O, Gursoy A (2015) Advances in template-based protein docking by utilizing interfaces towards completing structural interactome. *Curr Opin Struct Biol* 35:87–92
3. Keskin O, Nussinov R (2007) Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure* 15(3):341–354
4. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1996) Protein–protein interfaces: architectures and interactions in protein–protein interfaces and in protein cores. Their similarities and differences. *Crit Rev Biochem Mol Biol* 31(2):127–152
5. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1996) A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol* 260(4):604–620
6. Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A* 99(9):5896–5901
7. Mosca R, Ceol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. *Nat Methods* 10(1):47–53
8. Meyer MJ, Das J, Wang X, Yu H (2013) INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29(12):1577–1579
9. Hosur R, Xu J, Bienkowska J, Berger B (2011) iWRAP: an interface threading approach with application to prediction of cancer-related protein–protein interactions. *J Mol Biol* 405(5):1295–1310

10. Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins* 49(3):350–364
11. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A (2005) PRISM: protein interactions by structural matching. *Nucleic Acids Res* 33(Web Server issue):W331–W336
12. Tuncbag N, Gursoy A, Nussinov R, Keskin O (2011) Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* 6(9):1341–1354
13. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A (2014) PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Res* 42(Web Server issue):W285–W289
14. Cukuroglu E, Gursoy A, Nussinov R, Keskin O (2014) Non-redundant unique interface structures as templates for modeling protein interactions. *PLoS One* 9(1):e86738
15. Keskin O, Tsai CJ, Wolfson H, Nussinov R (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci* 13(4):1043–1055
16. Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O (2008) Architectures and functional coverage of protein–protein interfaces. *J Mol Biol* 381(3):785–802
17. Tuncbag N, Gursoy A, Keskin O (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* 25(12):1513–1520
18. Tuncbag N, Keskin O, Gursoy A (2010) HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res* 38(Web Server issue):W402–W406
19. Kuzu G, Gursoy A, Nussinov R, Keskin O (2013) Exploiting conformational ensembles in modeling protein–protein interactions on the proteome scale. *J Proteome Res* 12(6):2641–2653
20. Hubbard SJT, JM (1993) Naccess, Department of biochemistry and molecular biology University College, London.
21. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. *Proteins* 56(1):143–156
22. Mashiach E, Nussinov R, Wolfson HJ (2010) FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins* 78(6):1503–1519
23. Fiser A, Sali A (2003) Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol* 374:461–491
24. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 31(13):3381–3385
25. Pieper U et al (2014) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 42(Database issue):D336–D346
26. Nussinov R, Tsai CJ (2015) ‘Latent drivers’ expand the cancer mutational landscape. *Curr Opin Struct Biol* 32:25–32
27. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 43(Database issue):D789–D798
28. Stenson PD et al (2014) The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133(1):1–9
29. Landrum MJ et al (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42(Database issue):D980–D985
30. Forbes SA et al (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43(Database issue):D805–D811
31. Cerami E et al (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2(5):401–404
32. Acuner Ozbabacan SE, Gursoy A, Nussinov R, Keskin O (2014) The structural pathway of interleukin 1 (IL-1) initiated signaling reveals mechanisms of oncogenic mutations and SNPs in inflammation and cancer. *PLoS Comput Biol* 10(2):e1003470
33. Engin HB, Guney E, Keskin O, Oliva B, Gursoy A (2013) Integrating structure to protein–protein interaction networks that drive metastasis to brain and lung in breast cancer. *PLoS One* 8(11):e81035
34. Franceschini A et al (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41(Database issue):D808–D815
35. Tuncbag N, Kar G, Gursoy A, Keskin O, Nussinov R (2009) Towards inferring time dimensionality in protein–protein interaction

- networks by integrating structures: the p53 example. *Mol Biosyst* 5(12):1770–1778
36. Bos PD et al (2009) Genes that mediate breast cancer metastasis to the brain. *Nature* 459(7249):1005–1009
 37. Minn AJ et al (2005) Genes that mediate breast cancer metastasis to lung. *Nature* 436(7050):518–524
 38. Guney E, Oliva B (2012) Exploiting protein–protein interaction networks for genome-wide disease-gene prioritization. *PLoS One* 7(9):e43557
 39. Van Heyningen V, Yeyati PL (2004) Mechanisms of non-Mendelian inheritance in genetic disease. *Human molecular genetics* 13 Spec No 2:R225-233.
 40. Guven-Maiorov E et al (2015) The architecture of the TIR domain signalosome in the toll-like receptor-4 signaling pathway. *Sci Rep* 5:13128

NDEx: A Community Resource for Sharing and Publishing of Biological Networks

Rudolf T. Pillich, Jing Chen, Vladimir Rynkov, David Welker, and Dexter Pratt

Abstract

Networks are a powerful and flexible paradigm that facilitate communication and computation about interactions of any type, whether social, economic, or biological. NDEx, the Network Data Exchange, is an online commons to enable new modes of collaboration and publication using biological networks. NDEx creates an access point and interface to a broad range of networks, whether they express molecular interactions, curated relationships from literature, or the outputs of systematic analysis of big data. Research organizations can use NDEx as a distribution channel for networks they generate or curate. Developers of bioinformatic applications can store and query NDEx networks via a common programmatic interface. NDEx can also facilitate the integration of networks as data in electronic publications, thus making a step toward an ecosystem in which networks bearing data, hypotheses, and findings flow seamlessly between scientists.

Key words Biological networks, REST API, Java, Open source, BioPAX3, Big data, Pathway analysis, OpenBEL, Python, Cytoscape, Provenance, Cyberinfrastructure

1 Introduction

Networks are a precise and computable form in which biologists can express many kinds of information, including models of biological mechanisms, experimental facts, and relationships derived by systematic data analysis. When pathway diagrams evolved into repositories of small pathway networks [1, 2, 3], they became searchable resources and the basis for data interpretation and collaborative pathway editing applications [4]. In recent years, the development and progress in omics technologies have contributed to boosting the construction of networks inferred by the systematic processing of big data, providing an important avenue for interpretation and a counterpoint to literature curation [5, 6, 7]. In this view, it is necessary to create an infrastructure where the rapidly expanding corpus of biological network models created by

researchers can be stored, shared, discussed, reviewed, and used. This chapter describes NDEx, the Network Data Exchange, an online commons where scientists can store, share and publicly distribute biological networks as dynamic actionable data, and develop applications using them [8]. One of the goals of the NDEx Project is to create a home for network models that are currently available only as figures, tables, or supplementary information, such as networks produced via systematic mining and integration of large-scale molecular data. In doing this, the NDEx project does not compete with existing pathway and interaction databases, such as KEGG or Reactome; instead, NDEx provides a novel, common distribution channel for these efforts, preserving their identity and attribution rather than subsuming them.

By providing a flexible computable medium for biological knowledge, networks are also becoming a critical element for new models of scientific publication, in which data and its derivatives are as important as text [9]. The NDEx platform is intended to enable experimentation with novel forms of scientific review and discourse. NDEx networks are assigned stable, globally unique URIs and can be referenced by publications, by other networks, and by analytic applications.

NDEx aims to become the main hub for the development of new, lightweight applications and scripts capable of accessing and manipulating networks via the NDEx API, making it easy for scientists to develop novel network-based analyses. An example is CyNDEx, the NDEx Cytoscape App: CyNDEx enables users to access an NDEx server directly from Cytoscape and engage its full range of tools to analyze and transform any networks stored on NDEx. The NDEx project fosters the creation of new utilities and analytic tools that use NDEx via code examples, client libraries in multiple languages, developer documentation, and a strong social media outreach campaign to drive community awareness and engagement.

2 Materials

2.1 NDEx Sources

All NDEx sources are stored on GitHub in publicly accessible repositories under the **ndexbio** organization: <https://github.com/ndexbio>.

The following repositories support released NDEx software. For all other repositories under ndexbio, we advise you to consult with the NDEx team on their status before using. Some are in active development, while others may be obsolete or highly experimental.

2.2 NDEx Server

<https://github.com/ndexbio/ndex-rest>

<https://github.com/ndexbio/ndex-object-model>

<https://github.com/ndexbio/ndex-common>

- 2.3 NDEx Web App** <https://github.com/ndexbio/ndex-webapp>
- 2.4 NDEx Java Client** The NDEx Java Client is not an application; it is a library available for use by developers to create Java applications that access NDEx.
<https://github.com/ndexbio/ndex-java-client>
<https://github.com/ndexbio/ndex-object-model>
- 2.5 NDEx Python Client** The NDEx Python Client is a library facilitating access to networks on NDEx servers from Python. It provides convenient methods to find, query, and save networks while managing authentication.
<https://github.com/ndexbio/ndex-python-client>
- 2.6 CXIO** Java library is used for CX serialization and de-serialization by the NDEx server and the NDEx Java Client library.
<https://github.com/cytoscape/cxio>
- 2.7 CyNDEx Cytoscape App** <https://github.com/ndexbio/ndex-cytoscape-app>
<https://github.com/ndexbio/ndex-java-client>
<https://github.com/ndexbio/ndex-object-model>
- 2.8 NDEx Sync Copier Utility** <https://github.com/ndexbio/ndex-sync>
<https://github.com/ndexbio/ndex-java-client>
<https://github.com/ndexbio/ndex-object-model>

3 What Is NDEx

3.1 Comparison of NDEx to Other Network Resources

NDEx is an online resource to enable collaboration and publication using biological networks. It is a “commons,” a scientist-driven data exchange, where both individuals and organizations can share networks of any type, from pathway models and interaction maps in standard formats to a novel data-driven knowledge. This user-centric focus differentiates NDEx from the array of biological network resources currently available to biologists. In many cases NDEx is complementary to the missions of existing network resources, potentially playing roles as a novel distribution channel, a user content management component, or a source for staging of prepublication content. Most of these resources can be categorized either as **repositories** of network-structured information, as **analysis applications** that operate on input data (such as gene lists) via techniques that use one or more reference networks, or as **both**. “Repositories” in this context means resources where the network content is *managed by the organization maintaining the resource* and is therefore different from the structure of NDEx in which the *users manage the network content*. Some well-known

examples of repositories include KEGG (<http://www.genome.jp/kegg/pathway.html>), Pathway Commons (<http://www.pathway-commons.org/about/>), IntAct (<http://www.ebi.ac.uk/intact/>), and BioCyc (<http://biocyc.org/>). Many repositories also differ from NDEx because they use specific network formats and models of biology, in contrast to the NDEx strategy of supporting many formats in a common framework; one such example is the NCI/Nature—curated Pathway Interaction Database (<http://pid.nci.nih.gov/>) that only uses the BioPAX3 format.

Analysis applications using network resources include sites such as GeneMANIA (<http://www.genemania.org/>) and NCI DAVID (<https://david.ncifcrf.gov/>). Although NDEx provides some search and query operations that could be construed as “analysis,” its mission is not to perform biological analyses but instead to be a service that facilitates the creation of applications, both as a source of reference networks and as a place for users to store network-structured analysis results. A recent example of a network-oriented analysis application is the Network Portal (<http://networks.systemsbiology.net/>) by the Institute for Systems Biology, which “provides analysis and visualization tools for selected gene regulatory networks to aid researchers in biological discovery and hypothesis development.” Its design includes several features to promote data sharing and integration with other applications, but its primary focus is *analysis* using networks of transcriptional regulation.

WikiPathways [10] (<http://www.wikipathways.org/index.php/WikiPathways>) is a pioneering collaborative platform for the curation of biological pathways, a resource that shares the NDEx goal of facilitating scientific discourse by providing a platform for user-driven content. It differs, however, in that: (1) it is focused on pathway diagrams that are small and curated and in which the content may not be fully represented as a network, and (2) it employs the “Wiki” model of collaboration on a public document, different from the “Google Docs” approach of NDEx in which users manage the access to their networks. The role of NDEx in the context of collaborative environments such as WikiPathways could be as a “back end” resource to store and share the content created by the collaborators.

BioModels at EBI (<https://www.ebi.ac.uk/biomodels-main/>) is an example of a database of biological information that could be considered a network resource, but which is different from NDEx not only because the content is managed but also because it is specialized to a particular kind of biological data structure. BioModels is a “repository of computational models of biological processes,” serving as resource for the computational modeling community. Although there are forms of these computational models that can be expressed as networks (and which

NDEx may support at some point), BioModels presents these models in a comprehensive manner tailored to the needs of its user community of bioinformatic experts.

To conclude, NDEx provides a novel distribution strategy for organizations that maintain repositories, a new channel for their content to reach users and applications.

3.2 Available Online Resources

The following sections present a comprehensive but not exhaustive list of network repositories and selected examples of network-oriented analysis applications. The repositories include both those that are based on curated mechanistic information (“pathways”) and those that are focused on interaction data. PathGuide (<http://www.pathguide.org/>) was an invaluable resource for the preparation of this section.

3.2.1 Aggregators of Network Resources

Pathway Commons—<http://www.pathwaycommons.org/about/>

- “Pathway Commons is a network biology resource and acts as a convenient point of access to biological pathway information collected from public pathway databases, which you can search, visualize and download.”
- Aggregator of network repository data from many sources.
- Normalizes resources to BioPAX3.
- Distributes in SIF and BioPAX3 formats.

iRefIndex—<http://irefindex.org/wiki/index.php?title=iRefIndex>

- “Provides an index of protein interactions available in a number of primary interaction databases including BIND, BioGRID, CORUM, DIP, HPRD, InnateDB, IntAct, MatrixDB, MINT, MPact, MPIDB, MPPI and OPHID.”

3.2.2 Protein–Protein and Other Molecular Interaction Networks

Quoted sentences below and in Subheadings 3.2.4 and 3.2.5 are taken from the resources’ websites.

BioGRID—<http://thebiogrid.org/>

- “BioGRID is an interaction repository with data compiled through comprehensive curation efforts.”

CCSB Interactome—<http://interactome.dfci.harvard.edu/>

- A repository of experimentally derived protein interactions.

DIP—<http://dip.doe-mbi.ucla.edu/dip/Main.cgi>

- “The Database of Interacting Proteins (DIP™) catalogs experimentally determined interactions between proteins.”

IntAct—<http://www.ebi.ac.uk/intact/>

- A central, standards-compliant repository of molecular interactions, including protein–protein, protein–small molecule, and protein–nucleic acid interactions.

- IntAct provides both an open-source database system and analysis tools for molecular interaction data and acts as common curation platform for 11 molecular interaction databases.

NetPro—<http://www.molecularconnections.com/products/#netpro> “NetPro™ is a comprehensive database of Protein–Protein and Protein–Small molecules interaction, consisting of more than 320,000 interactions captured from more than 1500 abstracts, approximately 1600 published journals and more than 60,000 references.”

STRING—<http://string-db.org/>

- “STRING is a database of known and predicted protein interactions.
The interactions include direct (physical) and indirect (functional) associations.”

MINT—<http://mint.bio.uniroma2.it/mint/Welcome.do>

- “MINT focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators” and is now integrated with IntAct.

RBPDB—<http://rbpdb.ccb.utoronto.ca/>

- Repository of RNA–protein interactions.

BioLiP—<http://zhanglab.ccmb.med.umich.edu/BioLiP/>

- “BioLiP is a semi-manually curated database for high-quality, biologically relevant ligand-protein binding interactions.”

BindingDB—<http://www.bindingdb.org/bind/index.jsp>

- “BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules.”

TRANSFAC—<http://www.gene-regulation.com/index2>

- Commercial repository of gene regulation interactions, subset available for academic use.

iMEX—<http://www.imexconsortium.org/>

- “A non-redundant set of protein–protein interaction data from a broad taxonomic range of organisms.” iMEX also provides access to a selection of data curated by other resources such as IntAct, MINT, and DIP.

TAP Project—<http://tap.med.utoronto.ca/exttap/>

- “The Yeast TAP Project is aimed at elucidating the entire network of protein-protein interactions in a model eukaryotic organism, namely, the yeast *Saccharomyces cerevisiae*.”
- Repository derived from experimental data using tandem affinity purification (TAP).

3.2.3 Pathway Network Resources

NetPath—<http://www.netpath.org/>

- “NetPath’ is a manually curated resource of signal transduction pathways in humans.”

NCI-PID—<http://pid.nci.nih.gov/>

- The NCI/Nature—curated Pathway Interaction Database is a collection of “Biomolecular interactions and cellular processes assembled into authoritative human signaling pathways.”
- Officially retired as of December 31, 2015, the NCI-PID relies on NDEx as its primary distribution channel.

Reactome—<http://www.reactome.org/>

- “Reactome is a free, open-source, curated and peer reviewed pathway database whose goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology, and education.”

SignalLink Database—<http://signalink.org/>

- “An integrated resource to analyze signaling pathway cross-talks, transcription factors, miRNAs and regulatory enzymes.”

WikiPathways—<http://www.wikipathways.org/index.php/WikiPathways>

- “WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community.”
- Pathway diagrams are sometimes only partially computable, incorporating graphic elements with meaning apparent to the biologist but difficult for algorithms to interpret.

BioCyc—<http://biocyc.org/>

- “BioCyc is a collection of 5711 Pathway/Genome Databases (PGDBs), plus software tools for understanding their data.” It includes EcoCyc and MetaCyc and is focused on metabolism.

KEGG—<http://www.genome.jp/kegg/pathway.html>

- Repository of manually curated pathway and interaction networks and diagrams.

MANET—<http://manet.illinois.edu/aboutManet.php>

- “The Molecular Ancestry Network (MANET) database project traces evolution of protein architecture onto biomolecular networks.”

SMPDB—<http://smpdb.ca/>

- “An interactive, visual database containing more than 618 small molecule pathways found in humans.”
- Includes extensive, carefully formatted diagrams and exports in BioPAX3 and SBGN.

Atlas of Cancer Signaling Networks—<https://acsn.curie.fr>

- “ACSN is a pathway database and a web-based environment that contains a collection of interconnected cancer-related signaling network maps.”
- Uses SBGN created with cell designer and has a unique graphic interface.

UCSD Signaling Gateway—<http://www.signaling-gateway.org/molecule/>

- The UCSD signaling gateway molecule pages provide essential information on thousands of proteins involved in cellular signaling.

SPIKE—<http://www.cs.tau.ac.il/~spike/>

- “SPIKE is a database of highly curated human signaling pathways with an associated interactive software tool.”
- Incorporates information from other repositories in the curation process.

BIGG—<http://bigg.ucsd.edu/>

- “BiGG is a knowledgebase of Biochemically, Genetically and Genomically structured genome-scale metabolic network reconstructions.”

HumanNet—<http://www.functionalnet.org/humannet/>

- “A probabilistic functional gene network of 18,714 validated protein-encoding genes of *Homo sapiens* (by NCBI March 2007), constructed by a modified Bayesian integration of 21 types of ‘omics’ data from multiple organisms, with each data type weighted according to how well it links genes that are known to function together in *H. sapiens*.”

Ingenuity—<http://www.ingenuity.com/products/ipa>

- Large proprietary database of molecular interactions integrated with analysis tools.

Thomson Reuters Metabase—<http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metabase.html>

- Large, proprietary “manually curated database of mammalian biology and medicinal chemistry data.”

Pathway Studio—<http://www.elsevier.com/solutions/pathway-studio>

- Large proprietary database integrated with analysis tools.

3.2.4 Related Biological Repositories

BioModels—<https://www.ebi.ac.uk/biomodels-main/>

- BioModels Database is a “repository of computational models of biological processes.” Models described from literature are manually curated and enriched with cross-references.

The Cell Collective—<http://thecellcollective.org>

- Virtual cell models for simulations.
- Related to NDEx in that they also support a “crowdsourcing” strategy.

BioCarta—<http://www.genecarta.com>

Curated pathway diagrams. Not a network resource—only diagrams and gene lists are available and no computable connectivity.

3.2.5 Selected Examples of Network-Oriented Analysis

GeneMania—<http://www.genemania.org/>

- “GeneMANIA finds other genes that are related to a set of input genes, using a very large set of functional association data.”

Network Portal—<http://networks.systemsbiology.net/>

- “Provides analysis and visualization tools for selected gene regulatory networks to aid researchers in biological discovery and hypothesis development.”

DAVID—<https://david.ncifcrf.gov/>

- Gene set analysis enrichment scoring includes pathways.

MSigDB—<http://www.broadinstitute.org/gsea/msigdb/index.jsp>

- Gene set enrichment analysis.

GenomeSpace—<http://www.genomespace.org>

- “GenomeSpace is a cloud-based interoperability framework to support integrative genomics analysis through an easy-to-use Web interface.”
- Integration includes network-oriented tools.

Cytoscape—<http://www.cytoscape.org/>

- “An open source software platform for *visualizing* molecular interaction networks and biological pathways and *integrating* these networks with annotations, gene expression profiles and other state data.”

3.3 *Technical Description*

The following section will provide a short technical overview of the NDEx platform and its main components. For more detailed information on any of the following topics, please visit the documentation page on the NDEx website at www.ndexbio.org.

3.3.1 *Server and API*

The central component of the NDEx framework is the NDEx server, a database application accessed by a web-based **RE**lational **S**tate **T**ransfer **A**pplication **P**rogramming **I**nterface (**REST API**) [11], enabling its use from a wide variety of programming environments. Client libraries to facilitate the use of the API are available for Java, Python, and R. The network exchange format (CX) used in the NDEx API provides a semantic-neutral framework accommodating networks of many types while normalizing the treatment of controlled vocabularies, supporting references, provenance history, and property annotations of elements. The REST interface to the public NDEx server is deployed at <http://public.ndexbio.org/rest>.

3.3.2 *Web User Interface*

The main NDEx Web UI is a web application that enables users to search and browse networks, establish accounts, create or join groups, and upload, export, download, and administer networks. It provides simple visualization of small networks and query results from large networks and enables the user to review the provenance history of the network. The user can also edit network metadata, such as title or description. It is important to note that the Web UI relies only on the published NDEx API for access to an NDEx server, so it serves as an example for others to develop large NDEx-based applications. The Web UI for the public NDEx server is deployed at <http://public.ndexbio.org>.

3.3.3 *CX Interchange Format*

NDEx v1.3 supports network I/O operations where the network (or subnetwork in a query) is encoded in CX v1.0, the Cytoscape Cyberinfrastructure network interchange format. CX has been developed in conjunction with the Cytoscape group to be the future standard for network interchange by Cytoscape, NDEx, and applications and services in the Cytoscape Cyberinfrastructure (previous versions of NDEx used a JSON network format that will be supported for back compatibility but which is now deprecated).

The purpose of CX is to provide a format in which networks can be transmitted between diverse services (Fig. 1). It is designed for flexibility, modularity, and extensibility and as a message payload in common REST protocols. It enables applications to standardize on core aspects of networks, to coordinate on more specific standards within CX, and to ignore or omit irrelevant aspects. *It is not intended* as an optimized format for storage or for specific functionality in applications.

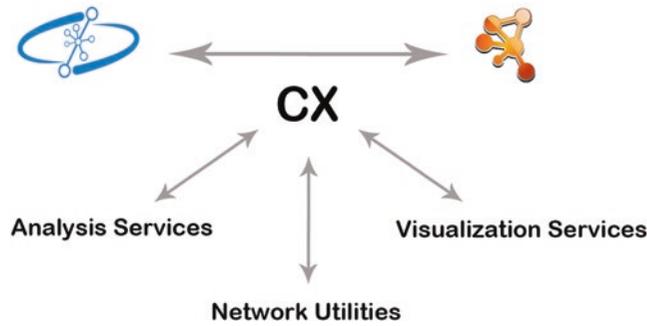


Fig. 1 The CX Interchange Format and its relationships with NDEx and Cytoscape in the context of the Cytoscape Cyberinfrastructure

The context in which CX will transmit networks is the Cytoscape Cyberinfrastructure (CI), a service-oriented architecture (SOA) implemented using REST protocols and primarily intended to facilitate computation with biological networks of many types. Within the CI framework, services may be created to store, query, visualize, lay out, and transform networks. These services may then be orchestrated by applications that implement bioinformatic analyses, ranging from custom user interfaces to interactively scripted applications in environments like IPython Notebook and to managed workflows in environments such as Taverna or Galaxy. An application such as the Cytoscape desktop application can participate both as a client of services and as a service for other applications.

Because of the wide diversity of network formats used in biology, a critical aspect of CX is that it provides straightforward strategies for lossless encoding of potentially any network. At the most basic level, this means that CX imposes very few restrictions: graphs can be cyclic or acyclic and edges are implicitly directed, but formats can choose annotation schemes to override this.

Semantically complex formats such as OWL, BioPAX, OpenBEL, SGML, or SBGN can be supported while at the same time enabling the expression of simple networks without undue overhead. CX does not, itself, make any commitment to a single “correct” model of biology or graphic markup scheme.

Finally, CX also addresses the critical requirement of enabling exchange of large networks while maintaining a low burden on both sending and receiving applications. Particularly, it enables a sending application to begin streaming a network without having the entire network in memory and allows the receiving application to begin processing a network before having received the entire network. Thus, the design of CX enables applications to reduce memory requirements for both the sender and receiver while also lowering delivery latencies.

3.3.4 Data Model

The CX network data model enables the transmission of networks with diverse semantics, uploaded from files in a variety of source formats, including SIF, XGMML, XBEL, and BioPAX3. Expressing these diverse network formats in CX provides users and application developers with consistent handling of nodes, edges, namespaces and identifiers, citations, properties associated with nodes and edges, and network provenance history. The CX data model does not, however, standardize the representation of biology in the networks that it stores. The meaning of the relationships indicated by edges or the classes indicated by the attributes of nodes in a network may conform to a rich standard such as BioPAX3 or OpenBEL, or they may have ad hoc meanings unique to the particular network. NDEx provides a common storage medium and access protocol for CX, facilitating the use of diverse networks by applications but not limiting the semantics that they may express.

The intent in the design of the CX network data model and in the NDEx utilities for loading specific network formats is to fully preserve the information content of networks: a network file in a given format imported to NDEx via CX should be equivalent (though not necessarily identical) to a network file output in a subsequent export operation using that format. The example in Fig. 2 shows one edge represented in the CX network data model.

Although NDEx provides both API and user interfaces to upload files in common formats (XGMML, XBEL, SIF, BioPAX3), the API also provides methods to create and query networks in CX, enabling researchers and developers to create and use networks with arbitrary semantics while still taking advantage of the common infrastructure supported by NDEx (Fig. 3). For example, researchers might experiment with novel representations of RNA–RNA and RNA–DNA interactions using CX aspects for controlled vocabularies, citations, or terminology definition by functional composition. When stored in NDEx, the resulting networks would benefit from NDEx-enabled applications that provide common functions such as basic visualization, indexing for search, or sharing and annotation. Specialized, modular applications (including ad hoc scripts) can then be constructed using the NDEx API to perform analyses and visualization that depend on the novel representation choices. This pattern of use is intended to foster experimentation with representations with rapid, straightforward sharing and discussion of the representational strategies and analytic consequences.

3.3.5 Network Search and Query

An Apache Solr (<http://lucene.apache.org/>) service running in a separate process provides the primary NDEx text search engine by indexing text strings to elements in the CX data model, including networks, users, groups, nodes, edges, citations, and supports.

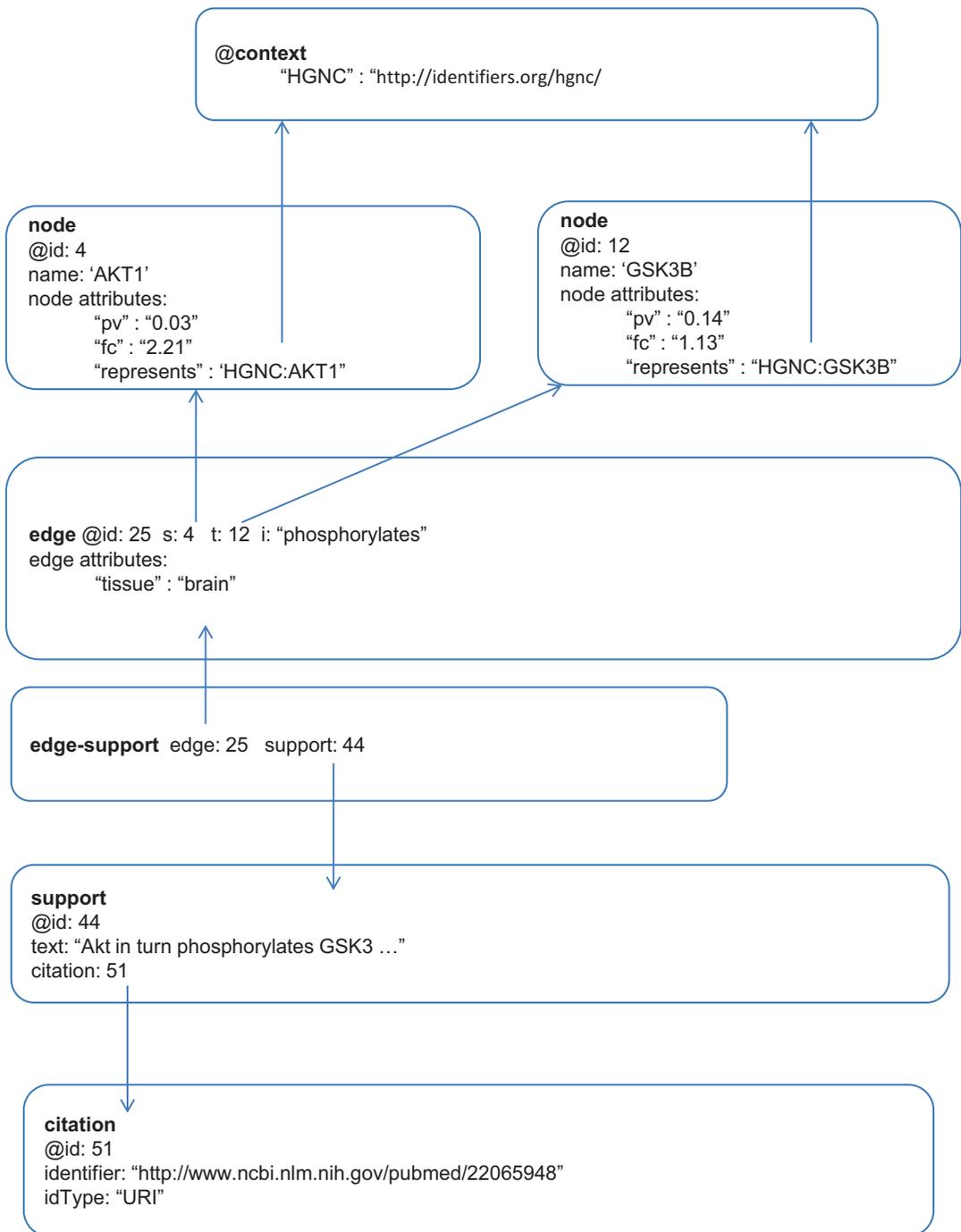


Fig. 2 Example of one network edge represented in the CX network data model. Each box in the diagram is a network aspect element, labeled with its aspect name and element id. Edge 25 connects nodes 4 and 12 by the s (source) and t (target) relationships. The meaning of edge 25 is indicated by its “interaction” attribute, “phosphorylates.” Nodes 4 and 12 are associated with terms “HGNC:AKT1” and “HGNC:GSK3B” using the “represents” attribute. These terms are linked by their prefixes to the HGNC-controlled vocabulary at a specific reference URI via the @context network aspect, indicating that the terms are standard human gene symbols. Both nodes have user-defined attributes “fc” and “pv” associated with them to record differential expression data mapped onto the network. Edge 25 has a user-defined property “tissue”: “brain” used by the authors to indicate the tissue context. Edge 25 is also annotated with an edge-support aspect element, justifying the relationship with evidence text from support 44. Support 44 is in turn derived from citation 51, the article from which the edge was curated

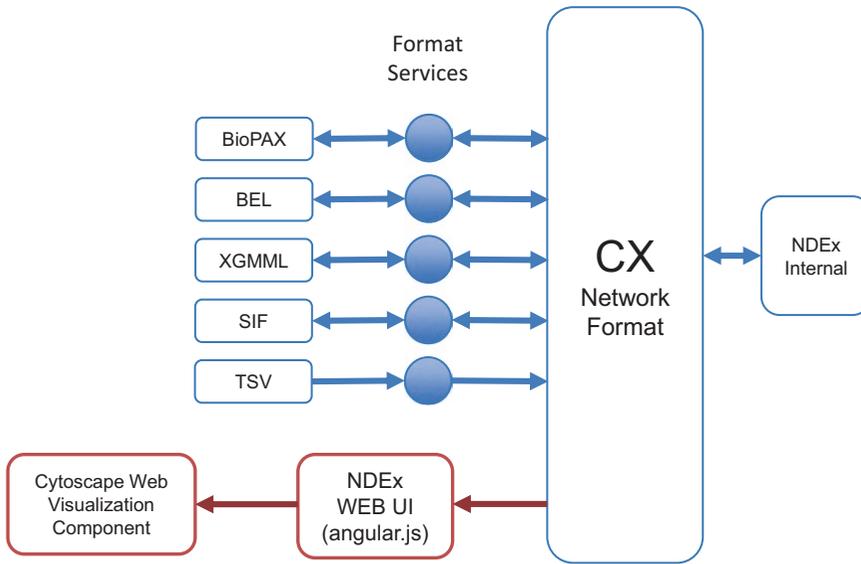


Fig. 3 NDEx v1.3 network format transformations

The Solr indexing enables search for networks by selected attributes. It also enables the identification of network nodes by their names and by controlled vocabulary terms, including terms in specialized CX aspects such as function terms.

The Solr service can run on the same machine instance as the NDEx server, a different machine instance, or even on a cluster. This flexibility provides an easy way to scale this aspect of NDEx, distributing the memory and processor requirements beyond a single machine. The NDEx architecture is explained in Fig. 4.

Because Solr is a standard, widely used search application, the configuration of the indexing of NDEx elements by text will involve less custom software and will be easier to modify, customize, and maintain. It is anticipated that this strategy will also result in improved performance, benefiting from Solr optimizations and future enhancements. The greater ease of customization will facilitate experimentation in indexing and improve application-specific optimizations. Any NDEx server, including the NDEx public server, can potentially extend the default indexing scheme.

The design for a default schema for attribute indexing using the Solr service is described in Table 1. The intent is to limit the default schema to a core set of network attributes that are justified/required in common use cases (such as the NDEx Web UI). The easy customization possible in this design enables a strategy of incremental testing and adoption of further attributes. This attribute schema does not describe the method for indexing networks based on the nodes that they contain. That method incorporates the node name and controlled vocabulary terms associated by CX aspects including nodeAttributes and functionTerms.

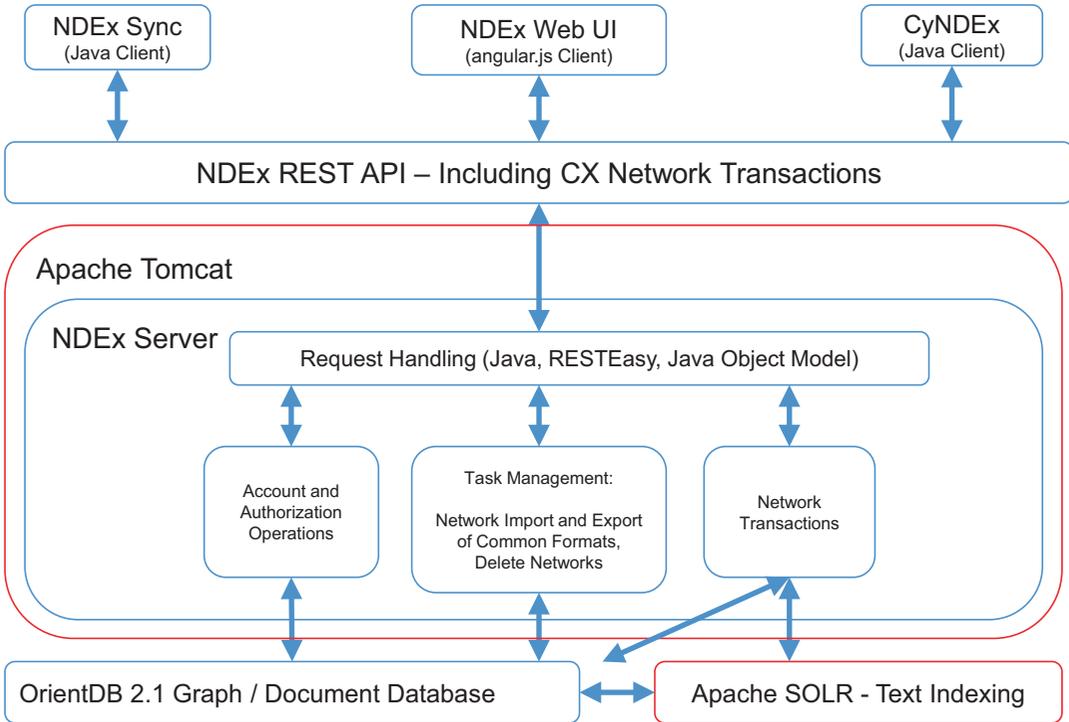


Fig. 4 NDEx v1.3 architecture

3.3.6 Usage Metrics

Although NDEx is a relatively young project, it has managed to gain a lot of momentum, thanks to its ambitious set of goals, collaborative efforts, and user-centric philosophy. Figure 5 shows some usage metrics in a simple graphical form. The future development and implementation of administrative tools will greatly enhance the spectrum of statistics that can be tracked, thus providing valuable material for analysis and targeted improvement.

4 A New Wave in the Publication Process

NDEx aims to provide a flexible infrastructure where users can simplify their workflow while promoting collaboration and interaction from the very earliest stages of the research process up until the final communication of results to the public.

4.1 Collaborating Before, During, and After a Publication

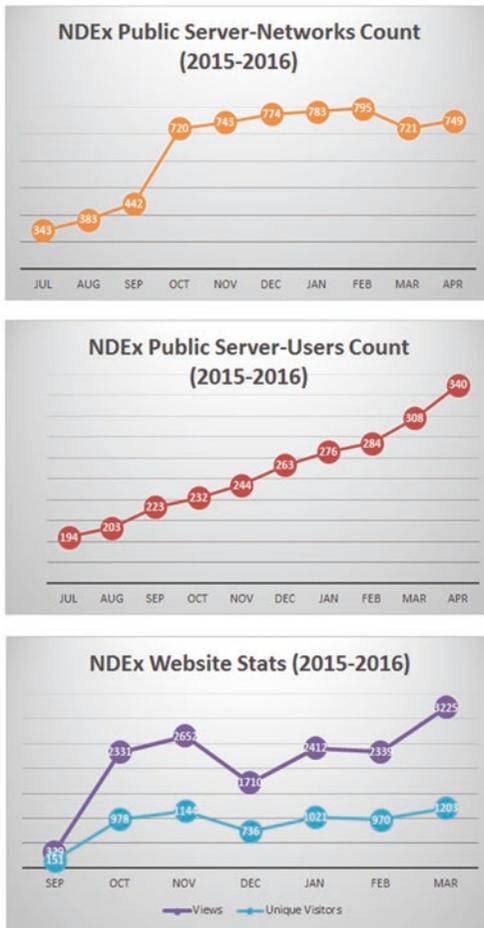
As a data commons, NDEx enables scientists and organizations that set up accounts on the NDEx server to upload and save networks and to create communities of users, much like Google+ Circles or LinkedIn groups. Users can manage access to their networks, making them private, public, or shared with selected users and community groups, similar to shared document systems such as Google Docs or Dropbox (Fig. 6). Networks stored in NDEx

Table 1

Default Solr indexing schema. Note that some attributes are highlighted in blue text. These are NDEX internal attributes and are not encoded in the CX networkAttributes aspect. In the case of attributes like “EdgeCount,” they are calculated values

Attribute	Description	Type
Name	Name of network	Freetext and string
Description	Description of network	Freetext
Version	Version of network	String
Organism	Description of organisms associated with the network	Freetext and string
Disease	Description of diseases associated with the network	Freetext and string
Tissue	Description of organs, tissue types, cell types, and cell lines associated with the network	Freetext and string
rightsHolder	Describes the holder of the rights to the network—same as dc:rightsHolder	Freetext and string
Rights	Description of the rights asserted by the rightsHolder—same as dc:rights. The indexing of this general field should aggregate not only the actual “rights” attribute but also the more specific rights attributes, such as copyright or license. In that way, network authors do not need to state information redundantly	Freetext and string
Copyright	Copyright asserted by the rightsHolder	Freetext and string
copyrightDate	The date of the copyright asserted by the rightsHolder	Date
licenseType	Identifies the type of license offered by the rightsHolder if a standard form is used	Freetext and string
publicRightsType	Identifies the network as having a public license of one of the following types: unrestricted/academicOnly/notForProfitOnly/researchOnly	String
Contributors	Individuals responsible for creating the content of the network—see dc:contributors	Freetext and string
creationTime	Time at which the network was created	Date
modificationTime	Time at which the network was last modified	Date
edgeCount	Number of edges in the network	Integer
nodeCount	Number of nodes in the network	Integer
Visibility	NDEX network visibility: public, private, discoverable	String
readOnly	NDEX	Boolean
Owner	Name of NDEX account owning the network	String

A



B

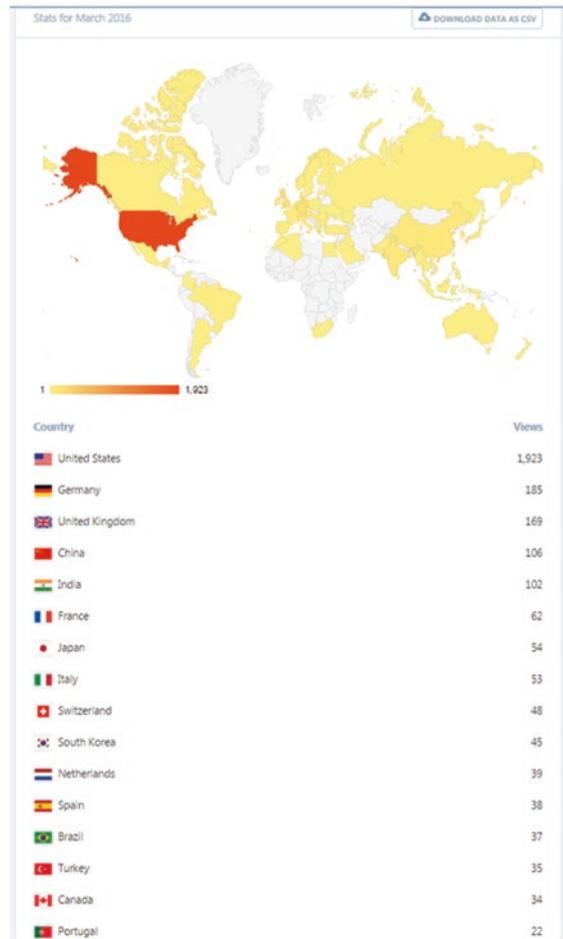


Fig. 5 Panel (a) shows the number of networks (*top*), registered users (*middle*), and numbers of unique visitors and views (*bottom*) for the NDEx public server starting from the NDEx v1.2 official release in July 2015. Panel (b) shows the geographical distribution of views in the month of March 2016

can be examined by reviewers during the publication process as live, interactive elements rather than static figures or complex data tables. As an example, the Ideker Laboratory at UC San Diego has recently started using NDEx to control access to networks referenced in submitted publications, thus simplifying and accelerating the review process.

To support publication of networks as data, it must be possible to unambiguously specify the identity of the network and trust that the content of a published network will remain constant. NDEx assigns every network a globally unique ID, a 128-bit UUID that distinguishes it from all others. A UUID reference to a network therefore refers to a specific data artifact, as maintained by its

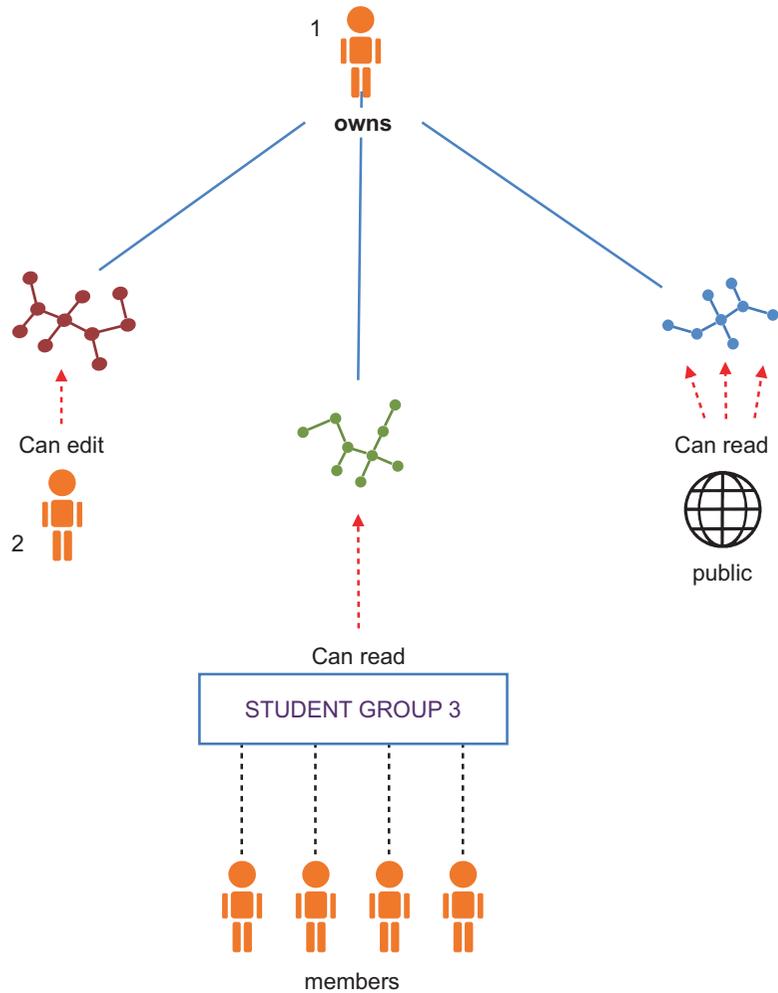


Fig. 6 Example of access control for networks in NDEx. User 1 owns the *red*, *green*, and *blue* networks. She makes the *blue* network public, allowing read access to anyone, including anonymous users; she shares the educational *green* network with the members of Student Group 3 for teaching purposes; finally, she shares the *red* experimental network directly with User 2 for research and collaboration

NDEx server. In addition, the owner of an NDEx network can set its status to be read only, preventing further edits. These features enable networks to be reliable, consistent references, suitable as inputs to further research.

Once published, NDEx seeks to facilitate the reuse of biological networks created by scientists as inputs to further experiment and analysis, providing structures that enable scientific reproducibility and new opportunities for attribution and citation. Given an analysis that uses information from public networks obtained from NDEx, it should be straightforward to reference those specific

networks and their authors. Moreover, any researcher wishing to validate the analysis should be able to access the exact networks used or, at a minimum, should know the identity and description of those networks. Finally, when a network is the result of a workflow in which other networks were inputs, it should be possible to access the history of events and sources for easy reproduction of results. At this stage, it becomes important to know how and when a network was created and which inputs and algorithms would be required to reproduce it. NDEx addresses these needs by including the “provenance history” with each network. The provenance history captures the workflow leading to the current network by describing prior events, networks, and other resources and grows as networks are created, modified, used, or copied. The provenance history is described in detail in the next paragraph and Fig. 7.

Finally, organizations that create and publish network content can use NDEx as a channel for distribution: networks from the NCI Pathway Interaction Database [12], the Pathway Commons [13], and the OpenBEL Consortium [14] are few examples of those available in NDEx. As an example, run a “user search” using the keyword “database”: NDEx will return all the user accounts of organizations maintaining publicly accessible network databases.

4.2 Network Provenance History

A network can represent assertions of biological relationships that are the results of experimental, analytic, or curation processes. Networks may in turn serve as inputs to further processes of analysis and model creation. If the workflow and dependencies on information sources are clearly documented, researchers may better understand the meaning of the relationships in the network and are better empowered if they wish to reproduce the analyses leading to the network. To achieve these goals, networks stored in NDEx include a provenance history aspect that can be accessed and managed via the NDEx API (Fig. 7).

For example, a network might be derived by an algorithm which finds subnetworks based on experimental data mapped to entities in a reference network; in this case the application performing the analysis should record the analysis event in the provenance history of the output network, including references or descriptions of the algorithm used, the input experimental data, and a description of the input reference network.

For robustness, the provenance history stores descriptions of “ancestor” networks and other information sources, not just links to those resources. This preserves the utility of the provenance history in situations in which some or all of the input information sources are unavailable or have been modified since they were used in the workflow. Researchers (or algorithms) can inspect the provenance history of the current network to address questions about the status of all of the inputs to the workflow. The NDEx network provenance history is similar in intent to synapse analytical provenance [15].

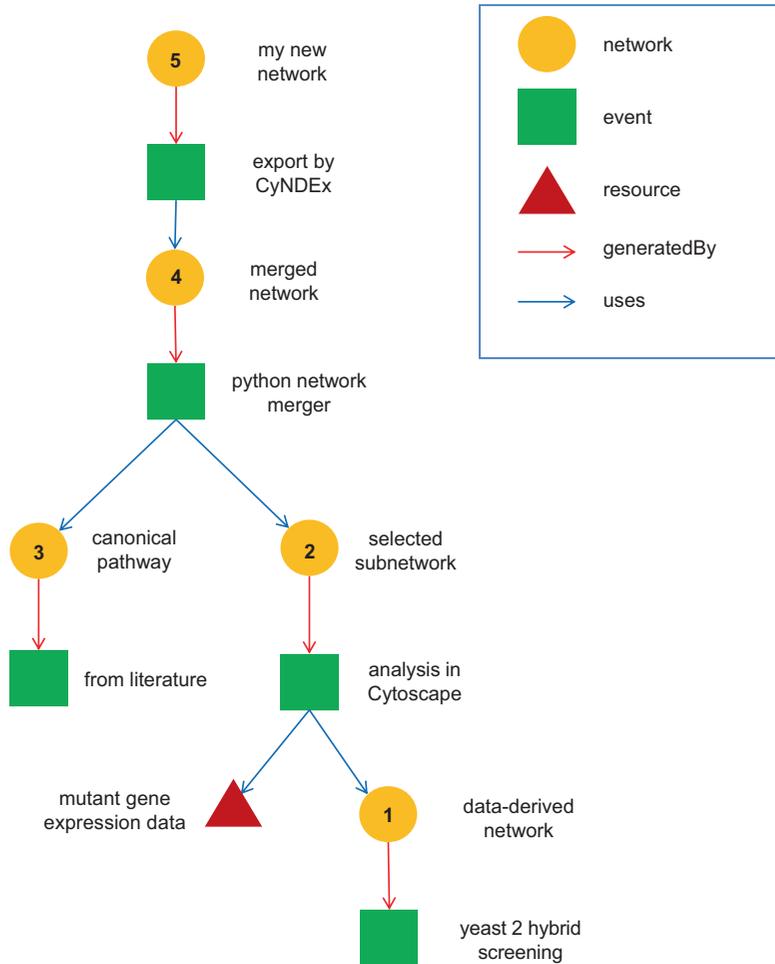


Fig. 7 A provenance history is a tree structure containing ProvenanceEntity and ProvenanceEvent objects. It is serialized as a JSON structure by the NDEX API. The root of the tree structure is a ProvenanceEntity object representing the current state of the network. Each ProvenanceEntity may have a single ProvenanceEvent object that represents the immediately prior event that produced the ProvenanceEntity, which is, in turn, linked to a network of ProvenanceEvent and ProvenanceEntity objects representing the workflow history that produced the current state of the network. The provenance history records significant events as networks are copied, modified, or created, incorporating snapshots of information about “ancestor” networks

The provenance history can be used to infer network equivalence, whether a given network stored in NDEX has the same content as another network or an external resource. This is valuable since in the general case, computing equivalence by algorithm may be computationally expensive or could require network format-specific knowledge. Any operation that modifies the network,

including changes to visibility or provenance, also changes the last modification date of the network.

Changes to network membership, what users have access to a particular network, do not modify the network itself and so do not change either the modification date or provenance history.

The standard fields in `ProvenanceEntity` and `ProvenanceEvent` objects correspond to relationships defined in the PROV-Ontology (PROV-O) [16]. Other property-value pairs can annotate these objects to provide more information about the entities and events. Any ad hoc pair of strings can be added as a property-value pair, and the properties used may be idiosyncratic to the recorded events and entities. However, the use of properties defined in the Dublin Core (DC) [17] metadata annotations and the Provenance, Authoring and Versioning (PAV) ontology [18] is preferred when applicable.

It is important to note the difference in the use of these ontologies in an NDEx provenance structure and the original intent. A `ProvenanceEntity` is a description of the referenced object, not the object itself. Therefore, a property such as `dc:title` that is asserted for a `ProvenanceEntity` refers to the original entity that the `ProvenanceEntity` represents. The provenance history references ancestor networks and other data sources but can also include self-contained descriptions of those objects that capture their state at the time they were used.

In a copy operation, an application/utility creates a new network (the target) that encodes the same content as an existing network (the source). In the resulting target provenance history, the root `ProvenanceEntity` represents the target, and the copy operation is represented as a `ProvenanceEvent` of type `COPY` in which the output is the root entity and the input is a `ProvenanceEntity` representing the source. The `ProvenanceEntity` representing the source and all of its prior entities and events are copied from the provenance history of the source. Information stored in the provenance history about the source is intended to reflect the state of the source at the time of the copy and should not be updated to reflect subsequent changes in the source. Information about the source stored in the provenance history is thereby preserved, regardless of whether the source is later modified or deleted.

In any case where the source network has the same UUID as the target, the `ProvenanceEvent` is an edit of some type. Because the event can have both `startingAt` and `endingAt` properties, the editing process can span an arbitrary amount of time. The application managing the editing process can therefore control the granularity of the provenance history. For example, an editing application could represent a long sequence of edits in a verbose chain of events and intermediate states, or it could simply keep updating the `endingAt` time as the edits continued. In both cases,

the resulting provenance history would be a valid representation of the workflow, although one would capture greater detail than the other.

In the case where a utility creates a network that has content equivalent or homologous to the source but described in a different identifier system (such as gene ids replaced with corresponding gene symbols), an additional resource describing the identifier mapping is typically involved. In this case, the mapping resource is also an input to the ProvenanceEvent, and it is appropriate to use the property pav:sourceAccessedAt to describe the relationship.

A modification operation in which the information in network A is augmented by information coming from network B, or where a new network is created from both A and B, is called a “merge.” This type of modification creates a branched provenance history (Fig. 7), where the ProvenanceEvent for the merge has two inputs, both network A and network B.

5 Applications and Services

Besides its roles as data commons and network publication channel, NDEx provides a flexible, programmatically accessible storage service that promotes modular software development and workflows; in this view, network output from one application can become input to another. NDEx is not meant to perform biological analysis and visualization itself but instead aims to enable the seamless interchange of networks between applications that are specifically designed for analytic purposes.

5.1 Communication Between NDEx Servers: The NDEx Sync Utility

While the public NDEx website is a resource to discover, access, share, or publish networks, there are situations in which it is better to deploy a separate copy of NDEx. For example, an NDEx server can be installed behind a firewall to handle cases where strong security is required, enabling storage of proprietary networks developed for the health science industry or those that incorporate patient information subject to privacy standards (e.g., the Health Insurance Portability and Accountability Act (HIPAA)). A private NDEx can also be deployed on local servers or on a scientist’s desktop computer for applications that store very large networks or perform frequent, expensive transactions (Fig. 8).

The information on a private NDEx server will, however, typically be most useful in conjunction with shared and public content. We have therefore created the NDEx Sync, a command-line utility that enables users to synchronize networks between different NDEx server instances, supporting cases where researchers wish to maintain local copies of public reference networks. NDEx Sync can be executed as periodic task, checking specified accounts and networks for changes on a “source” NDEx and then creating or updating networks on a “target” NDEx.

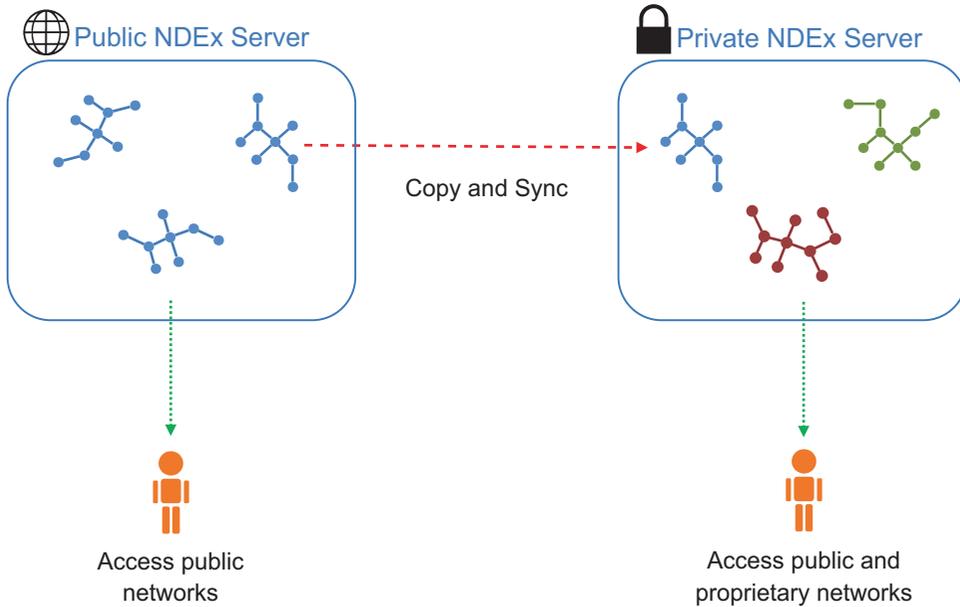


Fig. 8 Communication between NDEx servers. NDEx Sync can be used to copy and synchronize different NDEx servers, thus allowing access to both public and proprietary networks

5.1.1 How It Works

NDEx Sync is like a file-mirroring utility, but with an important difference: the copied networks are not exact duplicates of the source networks. Copied networks are assigned new UUIDs: every network stored in an NDEx server has a *globally unique* identifier and can be referenced by that identifier at its host NDEx. NDEx Sync updates (or creates, if necessary) the network’s provenance history, adding a “provenance event” that documents the copy process. The copied networks are therefore documented as *distinct entities*, copied at a specific time from a uniquely identified source. The provenance history provides a structure to document the events leading to the current state of a network. As already discussed, applications using NDEx are not required to maintain the provenance history for networks that they manipulate; however, it is strongly encouraged as a standard practice and will be supported by NDEx utilities.

For each *source network* that is selected as a candidate for copying, NDEx Sync examines the provenance history of *each network in the target account* to determine:

- Was this target network copied from the source network?
- Is the target out-of-date (i.e., the target “date last modified” is older than the source “date last modified”)?

The default behavior of NDEx Sync is that it will copy the source network to the target account if there is no copy of the source network in the target account or if the only copies present are out-of-date or have been modified.

5.1.2 Copy Plans

The working instructions for the NDEx Sync utility are defined by the user in a document called a “copy plan”; an example of a copy plan is shown in Fig. 9. The NDEx Sync copy plan specifies:

- An account and credentials for the source NDEx
- An account and credentials for the target NDEx
- The criteria to select networks on the source NDEx, which can be (1) a query to find networks matching search text, (2) a query to find networks administered by an account and matching search text, or (3) a list of network UUIDs
- The **updateTargetNetwork** parameter
- The **updateReadOnlyNetwork** parameter

5.1.3 Update of Networks by NDEx Sync

The default behavior of NDEx Sync is conservative, never overwriting or deleting any network in the target directory. This behavior can be overridden by the copy plan parameter **updateTargetNetwork**, specifying that NDEx Sync should update target networks that are identified as unmodified, out-of-date copies of the specified source networks. In an update, the target network keeps its UUID, but its contents are replaced by the contents of the source network, and the provenance history is handled in the same manner as in a default, non-update copy event. The updated network may be accessed by that UUID and any new request will obtain the updated content. Using NDEx Sync to update networks is only appropriate for situations in which the target network is intended as a cache of the source, where users want to obtain the *latest version* of the source content and where they do *not* expect the content of the network to be consistent over time.

5.1.4 Updates of Read-Only Networks

By default, updates will not be performed if the *target* network has read only => true. The **updateReadOnlyNetwork** configuration parameter in a copy plan overrides this behavior. This handles the case in which NDEx Sync is used to maintain a local copy of a remote resource and where the local copy is intended as a read-only, immutable reference. It is important to note that NDEx Sync can only update networks if the account specified by the *username* in the target element in the copy plan (*blue arrow* in Fig. 9) has administration privileges for the networks to be updated.

5.1.5 NDEx Sync and the CX Interchange Format

NDEx Sync uses the CX format for network copy and update operations in which both the source and target NDEx server are v1.3 or later. When a copy or update is performed in CX mode, each of the server transactions is a streamed operation. NDEx Sync does not need to instantiate the entire network in memory; instead, it simply operates by sending the output from the source stream to the input of the target stream. This reduces latency and, most importantly, the memory footprint on each machine involved in the process.

```

{
  "planType" : "QueryCopyPlan",
  "source" : {
    "route" : "http://public.ndexbio.org/rest",
    "username" : "user1",
    "password" : "pw1111"
  },
  "target" : {
    "route" : "http://localhost:8080/ndexbio-rest",
    "username" : "user2",
    "password" : "pw2222"
  },
  "queryString" : "*",
  "queryLimit" : "10000",
  "queryAccountName" : "user3",
  "updateTargetNetwork" : "true",
  "updateReadOnlyNetwork" : "false"
}

```



Fig. 9 Example of NDEx Sync copy plan. With this copy plan, User 2 will copy or update all networks (queryString = *) up to a maximum of 10,000 networks (queryLimit = 10,000) that are available in the User 3 account, from the source NDEx server (<http://public.ndexbio.org/rest>) to the User 2 account on the target NDEx server (<http://localhost:8080/ndexbio-rest>). *Blue arrow:* networks in the User 2 account will be updated only if User 2 has administration privileges on those networks

5.1.6 Command-Line Examples

The command is the same as the earlier version of NDEx Sync to run in CX mode:

```
> bash ndex-copier.sh/users/user12/my-copy-scripts
```

The “nocx” argument is used to run in the original mode. The original mode is *required* if one or both of the servers being accessed are running NDEx v1.2:

```
> bash ndex-copier.sh/users/user12/my-copy-scripts nocx
```

To run NDEx Sync from the MS Windows terminal, use the following commands:

```
> java -jar ndexbio-sync.jar my-copy-scripts
```

or

```
> java -jar ndexbio-sync.jar my-copy-scripts nocx
```

5.2 Communication Between NDEx and Cytoscape: The CyNDEx App

The Cytoscape desktop application and the wide variety of Cytoscape apps (plug-ins) provide a rich environment for analysis and manipulation of biological networks [19]. Cytoscape can also access NDEx via the REST API, enabling its users to search, query, import, and export networks. Under Cytoscape, a workflow might start by importing a transcriptional regulatory network from NDEx, after which the user could annotate the network with a differential mRNA expression dataset and process it to find subnetworks enriched for genes with significant changes in mRNA expression. The user could then export the subnetworks back to NDEx for review by collaborators or for use as inputs to further analyses.

Communication between NDEx and the Cytoscape desktop application is mediated by the CyNDEx App, and upcoming releases of Cytoscape are expected to incorporate its functionality into the main application, making NDEx networks immediately available to users [8]. The CyNDEx App is an example of an application in which the NDEx Java Client library uses the NDEx API to access NDEx servers.

CyNDEx allows Cytoscape users to connect to a selected NDEx server anonymously or via an existing account, search for networks, and then import a selected network or a subset of a large network as a query result. The imported networks are reduced to nodes, edges, and properties on those nodes and edges, in compliance with the Cytoscape table-oriented data model. Once the work in Cytoscape is done, users can export the modified network to NDEx and make it available to collaborators or the larger scientific community.

5.2.1 CyNDEx and the CX Interchange Format

CyNDEx v1.3 uses the CX format for network transfer from Cytoscape to NDEx. The use of CX enables the storage and retrieval of Cytoscape networks in NDEx without data loss, including the preservation of graphic markup and subnetwork structure. When a network is retrieved (imported) from NDEx, CyNDEx requests all of the CX aspects supported by Cytoscape: the *core* aspects and the *Cytoscape-specific* aspects. In contrast some aspects used in the NDEx Web UI such as citations or functionTerms are not supported by Cytoscape.

If the retrieved network was originally from Cytoscape, the combination of the core aspects and the Cytoscape-specific aspects restores the network in its entirety, bringing layout, graphic markup, and the structure of subnetworks and views. This Cytoscape data structure is displayed as a “network collection,” presenting views of its subnetworks that in some cases can be extremely simple, having only one view. CyNDEx provides two options for saving (exporting) networks to NDEx: either the entire network collection is saved as an NDEx network or a single view may be selected and saved.

If the retrieved (imported) network does not have any Cytoscape aspects, CyNDEx must create default structures in Cytoscape, building a simple network collection with a single view and default graphic style. If either the network collection or the single view is saved (exported) to NDEx, the new structures will also be stored.

Note that if a network with NDEx-specific aspects is retrieved and then saved back to NDEx as a new network, the NDEx-specific aspects *will not* be included in the new network. The only exception to this rule is the provenance history.

5.2.2 Network Update

Using the new NDEx API method **updateCXNetwork**, clients can update selected CX aspects and leave others untouched. CyNDEx can update a network stored in NDEx even though Cytoscape does not support all of the aspects in the original network. For example, a network using the NDEx-specific citation aspect can be imported into Cytoscape without its citations and then have Cartesian coordinates and other properties assigned. Those new properties, along with any other Cytoscape information, can be used to update the network in NDEx without changing the citations stored for that network. As discussed in the CX specification, it is possible to create networks that have inconsistencies between aspects managed by different applications, so applications are responsible for detecting and reconciling inconsistencies when they receive a network.

5.2.3 Provenance History

Although Cytoscape does not support the provenanceHistory CX aspect, CyNDEx preserves the provenance history of the networks it retrieves from NDEx when it creates new networks in NDEx based on these retrieved networks. Like the NDEx server, CyNDEx v1.3 relies on the **CXIO** Java IO library (*see* Materials) created by the Cytoscape team.

5.3 An Example of an NDEx Utility: The Python TSV Loader

Very often, network data consist of complex spreadsheets that can be hard to understand and challenging to interpret without carefully reading the publication in which they appear. The Cytoscape desktop application is capable of loading networks from Microsoft Excel spreadsheets or Tab-separated (TSV) text files; however, users are limited in their freedom to attach custom properties to the different network elements. In addition, there is no easy way to handle the citation information associated with nodes and edges. To overcome these limitations, we have created a versatile and fully customizable **TSV Loader** that takes advantage of the NDEx Python Client library (*see* Materials). The NDEx Python Client v2.0 was released as a PyPi module that can be installed via the PIP installation utility, and a detailed Python Client tutorial is available in the documentation section of the NDEx website at www.ndexbio.org.

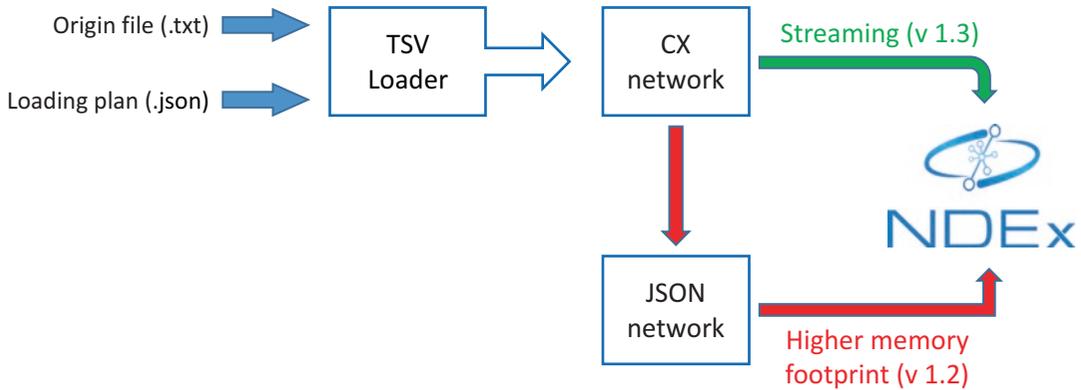


Fig. 10 NDEx TSV Loader. In NDEx v1.2, the TSV Loader creates a CX network, converts it to a JSON network, and then uploads it to NDEx. In v 1.3, the CX network will be directly uploaded to NDEx using the new CX streaming capability

5.3.1 Python TSV Loader

The TSV Loader has been developed to work with NDEx v1.2, the latest officially released version deployed to the NDEx public server at the time this chapter was written. However, we have taken into account the upcoming implementation of CX in NDEx v1.3 and made sure the TSV Loader was flexible enough to adapt to this new standard while being unaffected by the gradual phase out of the older platform.

In NDEx v1.2, the TSV Loader creates a CX network, transforms it in a JSON network, and then uploads it to the NDEx server (*red* path in Fig. 10). This requires the *full* network to be accommodated (stored) in the server’s memory, which can turn out to be a very expensive operation in cases where multiple users upload large networks at the same time.

In NDEx v1.3, the TSV Loader will create a CX network and upload it directly to the NDEx server using the CX streaming feature (*green* path in Fig. 10). This will dramatically reduce the memory footprint on the server’s side and guarantee better performance and efficiency. The next phase of development will wrap the TSV Loader utility in an intuitive UI and deploy it as a stand-alone service.

5.3.2 Origin File and Loading Plans

The TSV Loader is a command-line utility that, similarly to the NDEx Sync, uses JSON plans as templates for parsing networks for upload to NDEx. In addition to the loading plan, the TSV Loader also requires a tab-separated text file (origin file) that contains the actual network data. An example origin file and loading plan are shown in Fig. 11.

The loading plan consists of four main sections: source plan, target plan, edge plan, and citation plan (Fig. 11a). In each section, the user can specify what columns from the origin file (Fig. 11b)

a

```

1  {
2    "source_plan":
3    {
4      "context":
5      {
6        "uri": "http://identifiers.org/uniprot/",
7        "prefix": "uniprot"
8      },
9      "id_column": "Source ID",
10     "node_name_column": "Source name",
11     "property_columns": [""]
12   },
13   "target_plan":
14   {
15     "context":
16     {
17       "uri": "http://identifiers.org/uniprot/",
18       "prefix": "uniprot"
19     },
20     "id_column": "Target ID",
21     "node_name_column": "Target name",
22     "property_columns": [""]
23   },
24   "edge_plan":
25   {
26     "default_predicate": "interacts with",
27     "predicate_id_column": "",
28     "predicate_context": null,
29     "id_column": null,
30     "property_columns": ["DB entry"]
31   },
32   "citation_plan": {
33     "citation_id_columns": [
34     {
35       "id": "PMID",
36       "type": "pmid"
37     },
38     {
39       "id": "Article DOI",
40       "type": "DOI"
41     }
42   ],
43   "contributors_column": "Authors"
44 }
45 }

```

b

Source ID	Source name	Target ID	Target name	PMID	DB entry
Q2M3G0	ABCB5	Q5JNW4	PSMB9	8163024	http://www.uniprot.org/uniprot/Q5JNW4
P61221	ABCE1	Q70EL3	USP50	14715245	http://www.uniprot.org/uniprot/Q70EL3
P62736	ACTA2	Q13618	CUL3	10500095	http://www.uniprot.org/uniprot/Q13618
-----	-----	-----	-----	-----	-----

Fig. 11 (a) Example of a loading plan for the TSV Loader: users can fully customize the plan including namespaces, IDs, citation, and node/edge properties. **(b)** Sample origin file showing three protein–protein interactions with external links and PubMed IDs; the column names in this origin file are the same used in the loading plan shown in panel (a)

should be used to generate the network. The TSV Loader can easily handle citations, whether they are PubMed IDs or DOIs. Users are also free to append as many properties as they like to either nodes (source and/or target), edges, or both. In all those cases where a property is specified by a column in the origin file that contains a full URL (“DB entry” column in Fig. 11b), the TSV Loader will generate a clickable element, thus allowing seamless access to external resources right from within NDEx.

5.4 Outreach and Collaborations

The NDEx Project is carrying on several collaborations to develop and integrate analytic applications and utilities that will be essential elements of the Cytoscape Cyberinfrastructure. The CRAVAT/Mupit tool developed by the Karchin Lab [20] will use NDEx networks as back end for its enrichment analysis, while the Fraternali group at King’s College London is developing an analytic application for short loops in protein–protein interaction networks [21] that will use NDEx as source of networks, storage platform, and visualization. Additional collaborations with projects supported by the NCI ITCR program are still at a planning stage and will be starting in Q2 2016.

Since its inception, the NDEx Project has been developed keeping in mind that researchers might not be expert computational biologists; NDEx should be intuitive enough to be used by any researchers, from bioinformaticians to wet lab molecular biologists. In order to engage and educate users, the NDEx Project maintains a strong social media presence including a Twitter page (@NDExProject), a LinkedIn company page (<https://www.linkedin.com/company/ndex-project>), and a YouTube channel currently hosting only the NDEx Overview video but soon to be the reference point for video tutorials and user guides.

The NDEx Project is involved in the NCI Informatics Technology for Cancer Research Training and Outreach Workgroup (ITCR TOW) and in the J. Craig Venter Institute Network Special Interest Group (JCVI Network SIG) for the purpose of coordinating and improving outreach activities such as workshops, training sessions, and meetings.

In addition, the NDEx Project maintains an informational website (www.home.ndexbio.org) with news, blog, and the **NDEx Documentation**, both for users (manuals and tutorials) and developers (technical literature).

Acknowledgments

The NDEx Project is a joint effort of the Cytoscape Consortium and the Ideker Lab at the UC San Diego School of Medicine.

Primary financial support for The NDEx Project is from the National Cancer Institute (U24 CA-184427), F. Hoffmann-La Roche Ltd., Janssen Research and Development, LLC, and Pfizer, Inc.

NDEx is also supported by the National Resource for Network Biology (P41 GM103504).

References

- Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR et al (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res* 42:D472–D477
- Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30
- Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1998) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 27:29–34
- van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C (2008) Presenting and exploring biological pathways with PathVisio. *BMC Bioinf* 9:399
- Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3:140
- Califano A, Butte AJ, Friend S, Ideker T, Schadt E (2012) Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat Genet* 44:841–847
- Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. In *Nature methods* (United States). pp. 1108–1115
- Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, Stojmirovic A, Dobrin R, Braxenthaler M, Kuentzer J, Demchak B, Ideker T (2015) NDEx, the network data exchange. *Cell Systems* 1(4):302–305
- CyNetShare (2014) <http://cynetshare.ucsd.edu>
- Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C (2008) WikiPathways: pathway editing for the people. *PLoS Biol* 6:e184
- Fielding RT, Taylor RN (2002) Principled design of the modern Web architecture. *ACM Transactions on Internet Technology (TOIT)* 2:115–150
- Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) PID: the pathway interaction database. *Nucleic Acids Res* 37:D674–D679
- Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 39:D685–D690
- OpenBEL (2011) <http://www.openbel.org>
- Synapse Analytical Provenance (2015) <https://www.synapse.org>
- PROV-O (2013) PROV-O: The PROV Ontology <http://www.w3.org/TR/prov-o/>
- DublinCore (2012) Dublin Core Metadata Element Set, Version 1.1 <http://www.dublin-core.org/documents/dces/>
- Ciccarese P, Soiland-Reyes S, Belhajjame K, Gray AJG, Goble C, Clark T (2013) PAV ontology: provenance, authoring and versioning. *J Biomed Semantics* 4:37
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504
- Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R (2013) CRAVAT: cancer-related analysis of VAriants toolkit. *Bioinformatics* 29(5):647–648
- Chung SS, Pandini A, Annibale A, Coolen AC, Thomas NS, Fraternali F (2015) Bridging topological and functional information in protein interaction networks by short loops profiling. *Sci Rep* 5:8540

Bioinformatics Analysis of Functional Associations of PTMs

Pablo Minguez and Peer Bork

Abstract

Post-translational modifications (PTMs) are an important source of protein regulation; they fine-tune the function, localization, and interaction with other molecules of the majority of proteins and are partially responsible for their multifunctionality. Usually, proteins have several potential modification sites, and their patterns of occupancy are associated with certain functional states. These patterns imply cross talk among PTMs within and between proteins, the majority of which are still to be discovered. Several methods detect associations between PTMs; these have recently combined into a global resource, the PTMcode database, which contains already known and predicted functional associations between pairs of PTMs from more than 45,000 proteins in 19 eukaryotic species.

Key words Systems biology, Proteomics, Protein regulation, Post-translational modifications, Protein–protein interactions

1 Introduction

The cell is a very robust system where the final response to stimuli depends on many layers of regulation. In the last two decades, many new techniques have been developed to provide snapshots of genome and proteome regulation at different levels: transcription, posttranscription, translation, or posttranslation. These new types of experiments have changed partially the discovery workflow in science—now we may start to test a hypothesis with a lot of data and few assumptions, following a top-down strategy. Three main steps are required in this type of analysis: (1) filtering, mainly applying strong statistical controls in order to reduce the false positives rate; (2) annotation, where a biological meaning is superimposed on the statistics; and (3) integration, where results are merged with other levels of regulation to model cell behavior. Bioinformatics has come to help with the development of new algorithms, tools, and databases to address these three challenges.

For proteins, their role in a particular cell state is partially controlled by means of the addition of small moieties called post-translational modifications (PTMs). There are many PTM types described that are amino acid specific; their number, position, and combination present at a particular moment determine the final state of a protein. Mass spectrometry (MS) technology is able to explore the modification status of the whole proteome with a particular type of PTM at once. This technique requires an enrichment of these moieties in order to increase their detection threshold, which results in false positives as a side effect. An increasing number of experiments are now available that report mostly phosphorylation sites but also the occupancy of other PTM types such as acetylations, ubiquitinations, glycosylations, and others, of nearly the whole proteome under different conditions. Under this scenario there are two big challenges to address in order to translate the huge amount of data available into reliable information about the regulation of specific proteins. First, we should try to discriminate among all PTMs reported to focus on those with biological relevance. Several databases work to gather this type of information, measuring the conservation of the modified amino acids as a proxy for the prevalence of the PTM over evolution [1, 2], calculating the accessible surface area [3], or mapping the PTMs onto more or less complete maps of the regulatory elements of the proteins [1–5] and even onto their secondary [3] and tertiary structures [1, 3, 6]. The second challenge is to elucidate the possible cross-regulatory effects of PTM combinations under specific conditions. There are hundreds of described examples of PTM cross talk in the literature [7, 8] (*see Note 1*), and it is postulated that the function, localization, and interactions of most proteins depend partially on their modification pattern [9]. However, the search for their cross-regulatory effects is still an unexplored field. There are a few systematic efforts that go beyond the basic annotation of the experiment where the modifications were reported [4] (e.g., the cell cycle phase where they are present [2]) to more detailed information; for example, the downstream effects of the modifications [10] and their interaction with other PTMs [1] have been manually annotated, and text-mining tools have been used to report the functional processes in which the modifications are involved [11]. Although these approaches provide very accurate and valuable information, they can only be applied to low-throughput experiments (LTEs), while the majority of the PTMs reported come from high-throughput experiments (HTEs). The PTMcode database complements this information with several prediction methods to annotate pairs of PTMs as being functionally linked. To the best of our knowledge, PTMcode is the only resource that provides predictions of post-translational regulation of proteins based on the functional associations among their PTMs. In addition, PTMcode also provides predicted cross talk between

PTMs in interacting proteins that might regulate their binding, placing PTMcode as a bridge between protein-protein interactions (PPIs) and PTM databases.

Herein, we describe in detail how to explore functional associations between protein modifications using the PTMcode database.

2 Materials

Two types of users can profit from the data produced by the PTMcode database: those interested in the post-translational regulation of particular proteins or protein families [12, 13] and those interested in performing further high-throughput computational analyses on the PTMs, either using our curated dataset of PTMs in one or many species [14–16] or applying other algorithms to the predicted functionally associated PTMs in order to contribute to the deciphering of the global “PTMcode” [17, 18] (*see Note 2*). For both type of tasks, the PTMcode database has implemented ways to explore and download the data.

PTMcode is freely accessible at <http://ptmcode.embl.de> and requires no more than a modern browser installed in a standard computer. In order to access some PTMcode features, users should allow java applets to be run. Please check if your favorite browser is supported here (https://java.com/en/download/help/enable_browser.xml).

3 Methods

3.1 PTMs and PPIs Are the Main Sources of Protein Post-translational Regulation

Two interconnected processes regulate final protein function and localization: (1) the interaction with other molecules, mainly other proteins to form transient or stable complexes, and (2) the addition of PTMs [19] that may regulate the protein’s binding activity. Many resources are dedicated to gather and curate these two types of events. The PTMcode database is not a substitute for PTM databases [2, 11, 20] nor the repositories that compile or predict PPIs [21–23]; instead it incorporates both types of regulation in a unique resource in order to provide a complete picture of the post-translational regulation of eukaryotic proteins. Thus, PTMcode compiles in its second release:

1. Post-translational modifications from six databases [2, 3, 11, 20, 22, 24] and nine proteome-wide experiments [25–33] summing up 316,546 experimentally verified PTMs of 69 different types. The PTMs are mapped onto sequences extracted from the eggNOG database [34], which chooses the largest transcript as representative of the protein. To avoid spurious

mapping due to different sequence or transcript versions, PTMcode checks that all PTMs for a particular protein coming from the same source modify the type of amino acid reported. If even one PTM is mapped onto the incorrect residue, we assume that the protein sequence from the source is not the same as the one in our database, and all PTMs coming from that source for that protein are discarded. This methodology permits us to build a consistent and accurate dataset of protein modifications for 45,361 proteins from 19 eukaryotes.

2. Protein-protein interactions extracted from the STRING database [35]. STRING compiles and scores known and predicted protein-protein associations based on several types of evidence. PTMcode collects high-confidence PPIs (score >0.700) that are based on experimental evidence of physical binding (“experiments” evidence type). In total, PTMcode includes 221,268 PPIs from the 19 eukaryotes.

3.2 Non-Experimentally Verified PTMs

In order to understand the modification landscape of the proteome, the study of individual proteins by means of LTEs, although very accurate, is clearly insufficient. We need HTEs to cover as many species, proteins, PTM types, and conditions as possible. Despite the efforts of the community, this goal is still far from being accomplished as only a few species have been subject to this type of screening. Thus, many different tools have been developed to predict modifications (<http://www.cbs.dtu.dk/databases/PTMpredictions/>).

The PTMcode database annotates a new category of PTMs that complements the information provided by experimentally verified modifications—the so-called “propagated PTMs.” The principle behind them is that conservation of an amino acid over evolution is a proxy for the conservation of its function and so for its modifications [36–38].

Propagated PTMs are assigned using orthologous groups (OGs) from the eggNOG database. eggNOG builds OGs for proteins from thousands of species and organizes them in levels of inclusive taxa. For every protein, PTMcode selects the OG of the oldest eukaryotic level in which it is included and spreads the annotation of its experimentally verified PTMs across the conserved residues in a multiple sequence alignment (MSA) of all proteins in that group (Fig. 1). This naive exercise allows us to disseminate modifications from one species to others. We evaluated that 22.7 % of the experimentally verified human phosphoserines align with a known phosphorylation site in another species (15 % is the random expectation) [1], which is surely an underestimate as HTEs have been performed on only a few other species. Thus, the PTMcode database maps over 1,30,000 non-verified PTMs in ~130,000 proteins and provides modification patterns for the proteomes of species that have not been subjects of HTEs.

opiNOG06286 Orthologous Group
(part of the Multiple Sequence
Alignment)



Fig. 1 Schema for PTM propagation. The SOX2 protein has two sources of PTMs coming from two HTEs performed in mouse and human (sequences highlighted with a *red rectangle*). The experimentally verified PTMs are mapped into the MSA of the OG, and the conserved amino acids in the columns with annotated PTMs are marked as “propagated PTMs”

To show the impact of this type of novel annotation, nine species had, as a result of our “PTM collection pipeline,” less than 500 “real” modifications, but by including propagated PTMs, these numbers increased more than 250-fold to a level comparable with species with HTE data.

PTMcode predicts functional associations for propagated PTMs among themselves and with experimentally verified modifications. However, the propagated PTMs have to be considered potential PTMs and should be interpreted with more care than the ones found in experiments.

3.3 Channels for the Prediction of Functional Associations Between PTMs

There are many possible ways in which two particular PTMs might be functionally associated. For instance, they could be part of a molecular switch that controls protein function and/or localization [39], they could constitute a series of consecutive modifications [40], or they could contribute to the same final outcome of the protein even though they are added at different times and in different cell compartment (e.g., PTMs as a signal for protein transport).

In order to catch this wide variety of regulatory events, the PTMcode database implements five independent channels to predict the functional association between PTMs within the same protein; some of these are also applied to PTMs between interacting proteins. Below we describe these five channels in the context of PTMs within the same protein; for the association of PTMs between interacting proteins, see Subheading 3.7.

3.3.1 Coevolution Channel

The coevolution of two protein residues has been widely used as a proxy for their functional connection [41]. One of the most popular algorithms to address this concept is mutual information (MI) [42]. When applied to the MSA of a group of orthologous proteins, MI can estimate the coevolution of two residues (two columns in the MSA) by measuring the accuracy of predicting the amino acid

present in one position knowing the identity of the amino acid in the second position. To evaluate the functional association of two PTMs, PTMcode uses a slightly modified version of MI to penalize anticorrelation of residues (residues that have an opposite pattern of coevolution).

As discussed in [43], the signal coming from the MI evaluation must be compared to a background distribution of MI values to avoid spurious correlations due to phylogenetic influences (closely related species in the MSA) and small sample size (few species in the MSA). Some background distributions have been already proposed based on label randomization in the MSA [44] or on the set of MI values of all pairs of residues in the MSA [43]. PTMcode uses very strict criteria for its background distribution, using the MI values from non-modified residues in the MSA of the same type of amino acid and located in similar protein regions (ordered or disordered) as the two modified residues under evaluation. Pairs of PTMs with an MI value higher than 95 % of the background distribution are classified as coevolving. For residues lacking enough variability in the MSA column (very conserved or not conserved) to be able to compute MI values, we calculate the ratio of the conserved site in a MSA position to the total proteins and compare it to the distribution of the non-modified sites with the same limitations taken as background. Again pairs with a ratio above 95 % of the background distribution are selected as coevolving.

The coevolution channel is designed to extract a wide range of regulatory relationships as the underlying mechanisms might be very different and are not included in the definition of the algorithm.

3.3.2 Structural Distance Channel

The atomic proximity of two amino acids in the 3D structure of a protein is a widely accepted proof of the residues' functional association. Indeed protein contact maps representing matrices of all-against-all residue distances within a protein can be used to reconstruct its 3D structure [45]. The PTMcode database uses available protein 3D structures from the Protein Data Bank [46] to measure the distance between C α -C α atoms of all pairs of modified residues within the model. Although a threshold of 6–12 Å is usually accepted to determine contact between residues, we wanted to be more strict, so we calculated the threshold value based on known cases of PTMs that have a physical interaction. In total, we could measure 12 pairs of residues having this type of association and set as an optimal value for physical contact their average distance (4.69 Å).

The limitations of this channel are due to the availability of 3D protein structures and the mapping of the residues from sequence coordinates to positions in the structure.

3.3.3 *Same Residue Channel*

A very specific case of PTM cross talk is the direct competition of two types of modifications for the same protein residue [47]. There are two significant examples of this type of cross talk, the yin-yang molecular switches [48] where the same serines or threonines are modified with a phosphorylation or an O-linked glycosylation, co-regulating protein function and localization, and lysines that can be acetylated, SUMOylated, ubiquitinated, or methylated to produce different outcomes (e.g., in histone tails [49]).

PTMcode collects these events by checking the different PTM types that modify the same protein residue.

3.3.4 *Manual Annotation Channel*

In addition to predictions, PTMcode stores manually annotated PTM crosstalk events extracted from published papers. In these cases, a description of the interaction of both PTMs is provided.

3.3.5 *PTM Hotspot Channel*

From the analysis of the post-translational regulation of well-studied proteins such as the TP53 oncogene, we have learned that there are certain protein regions with an accumulation of PTMs [50] that act as regulatory centers (PTM hotspots). This concept was extended by Beltrao et al. [51] to many other eukaryotic proteins. PTMcode identifies PTM hotspots following the definition of Beltrao et al. and presents them within its complete framework of regulatory events. For each modified residues in a protein, we count the number of PTMs in a window of 31 amino acids (15 downstream and 15 upstream) and compare them using a Fisher exact test to the number of modifications in the whole protein. *P*-values are adjusted by false discovery rate and overlapping regions are collapsed.

3.4 *The PTMcode Home Page*

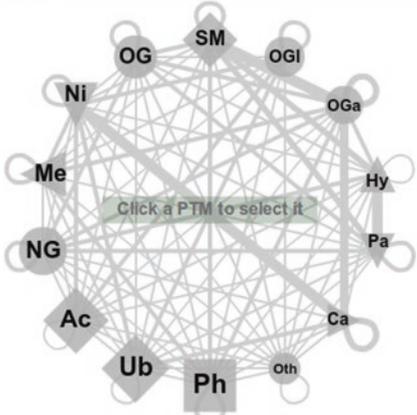
The users' entry page (Fig. 2) is divided into two panels:

1. The left panel (Fig. 2a) is dedicated to the exploration of particular combinations of PTM types. It is an option implemented for scientists that are interested in a particular type of cross talk, e.g., phosphorylations and O-linked glycosylations in yin-yang sites [52, 53] or phosphorylation linked to ubiquitination as a signal for degradation [7, 54]. From the browser wheel, users may select two types of PTMs and a table of all instances of known and predicted functional associations are displayed below. Two tabs divide the table into associations within the same protein and those regulating two interacting proteins.
2. The right panel (Fig. 2b) is designed to allow users to explore the predicted regulation of particular proteins. Again, two tabs separate the options to explore the regulation of a protein of interest or the regulation of its interactions with other proteins. The input is either a protein name or sequence (several examples are provided), and the search can be restricted to specific protein regions, residues, or PTM types of interest.

Known and predicted PTM functional associations

PTMCode is a resource of known and predicted functional associations between protein post-translational modifications (PTMs) within and between interacting proteins. It currently contains 316,546 modified sites from 69 different PTM types which are also propagated through orthologs between 19 different eukaryotic species. A total of 1.6 million sites and 17 million functional associations more than 100,000 proteins can currently be explored.

Browse PTMs



A network diagram showing relationships between PTM types. Nodes are labeled with PTM abbreviations: OG, SM, OGI, OGa, Ni, Me, NG, Ac, Ub, Ph, Oth, Ca, Pa, Hy. A central text box says "Click a PTM to select it".

Explore a single protein

Explore a protein pair

Use either a protein name/identifier or its full amino acid sequence. You can limit the PTMs reported to any specific residue or protein region of interest. If you are interested only in one or several PTM types, select them in through the **PTMs of interest** pulldown box.

Protein sequence or identifier

Examples: #1 #2 #3 #4

Options

Residue: eg. Y251

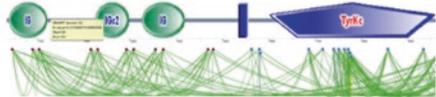
Protein region: **to** eg. 50 to 300

PTMs of interest: 14 of 14 options

Species: Any

Explore protein

Reset



A schematic diagram of a protein with several PTM sites indicated by colored circles (green, blue, red) and a blue arrow pointing to a specific site. The protein is shown as a blue arrow with a tail.

Fig. 2 The PTMcode home page

3.5 Exploring a Particular Protein

From the home page (<http://ptmcode.embl.de>), PTMcode provides several options to search for a protein of interest. Users may enter either the sequence (only exact matches are reported) or a protein identifier. In the case of protein IDs, PTMcode uses the protein ID dictionary from the STRING database, which has cross links between IDs from the major protein and gene resources. This dictionary does not report only synonymous IDs for the same protein but other cross links that help the users to identify the correct protein from the melting pot that represents the world of genes and protein names. If there is any source of conflict in the name provided, we redirect the user to a disambiguation page where the correct entry can be selected. If, in spite of the facilities implemented to find the desired protein, there is an unclear outcome or the protein is not found, we suggest searching for the protein ID in the Ensembl database (<http://www.ensembl.org/>) and, following the links to proteins and transcripts, getting either the ID or the sequence of the largest transcript associated with the input ID. Be aware that the protein sequence is our ultimate unique identification for a particular protein.

Users may also restrict the search to their favorite PTM types, known modified residues, or particular protein regions.

Once a protein is selected, PTMcode directs the user to its entry page that is divided into three panels (Fig. 3):

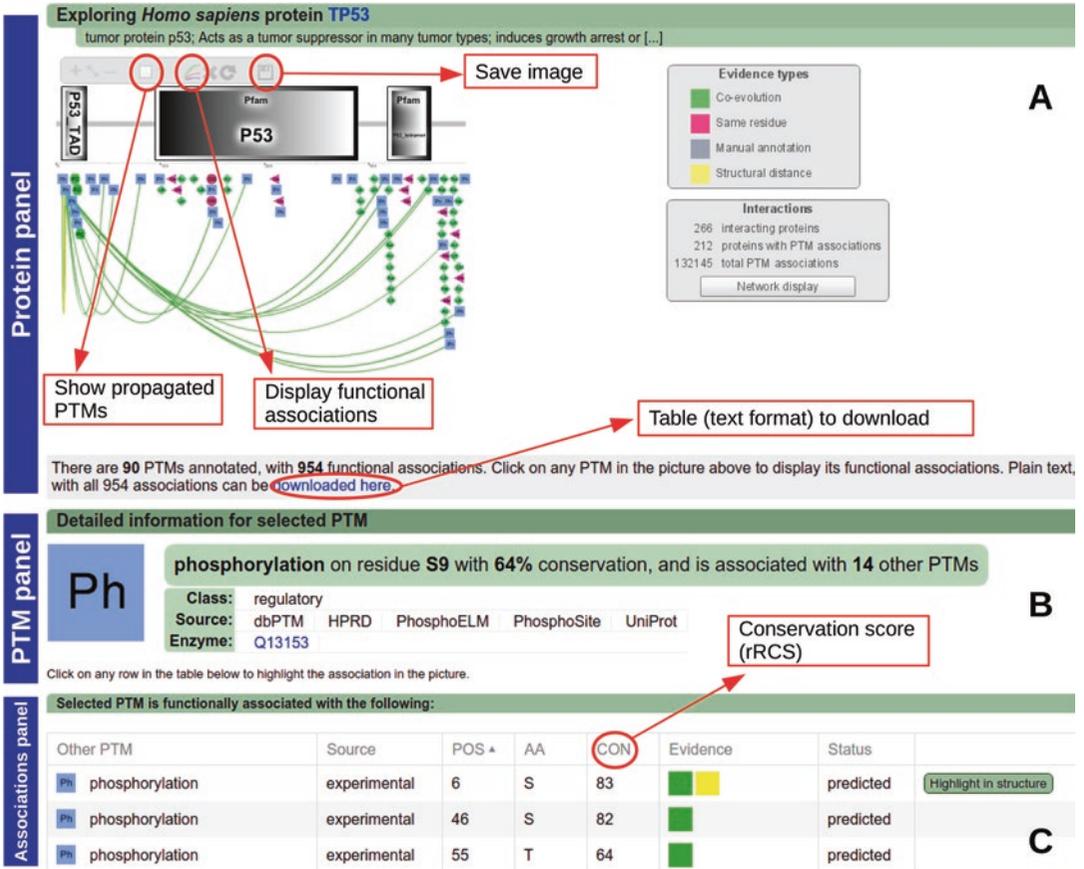


Fig. 3 PTMcode Results page. The three panels are shown for the TP53 human protein. In the “protein panel” (a), users may choose to show the “propagated” PTMs and display all functional associations or those related to one particular PTM or evidence type. They also can download the set of known and predicted associations either as an image or in text format. In the “PTM panel” (b) detailed information about the selected modification is shown. From the “associations panel” (c), the details of every PTM functionally associated with the selected modification are displayed including their rRCS (highlighted in red) and the evidence channels supporting the prediction

1. The protein panel. An interactive framework where an image of the protein is displayed (Fig. 3a) showing along the protein coordinates: (1) the protein globular domains annotated by the SMART database [55], (2) the PTMs, and (3) the hotspot regions. From this panel, users may explore the information about the domains (linked to SMART), PTMs, and their functional associations with other modifications. A zoom facility allows viewing of the details of particular protein regions and a checkbox permits inclusion of the “propagated PTMs” in the display. They are shown with a red border to distinguish them from the experimentally verified modifications. When the user clicks on their favorite PTM, the rest of the panels refresh in order to show the knowledge for that particular modification.

2. The PTM panel. Under the header “detailed information for selected PTM,” a self-refreshing table appears each time a PTM is clicked in the “protein panel” (Fig. 3b). In here, some of the features of the selected PTM are displayed:
 - (a) The type of modification.
 - (b) A conservation score that reflects how the residue is conserved across orthologous proteins.
 - (c) The number of functional associations predicted.
 - (d) A general classification of the possible function of the PTM based on the PTM type: categories include “regulatory” (involved in regulation of protein function), “stabilizing” (required for conformational purposes), and “uncharacterized” (with unknown or unclear function).
 - (e) The source from which we obtained the modification.
 - (f) The enzyme that performs the modification (if annotated in the sources at the time of downloading the data).

The information provided by this panel is complemented by a pop-up box that appears when the mouse is over a PTM in the “protein panel.”

3. The PTM associations panel. A user-sortable table that lists the modifications that were predicted to be functionally associated with the selected PTM (Fig. 3c). The table provides (a) the modification type, (b) the source (in release 2 all are experimental), (c) the amino acid and position in the sequence, (d) the conservation score, (e) the set of evidence (channels) that supports this association, and (f) the status, either predicted or known.

The different questions that can be answered from these three panels are discussed in detail within the following subheadings.

3.6 How to Assess if a PTM Is Biologically Relevant

One of the hot topics in the field of protein PTMs is the discrimination between functional and nonfunctional modifications. This information is especially relevant for phosphorylation sites as from the early days of MS proteome-wide experiments; it has been postulated that some phosphorylations might result from promiscuous kinase activity [56]. In addition, it is possible that in some cases the conservation of phosphorylation events might be at the kinase-substrate level and not at the site level [38], especially in disordered regions. Still, phosphorylation sites have been found to be generally more conserved than both their flanking regions [57] and non-modified serines, threonines, and tyrosines in ordered and disordered regions [9]. Thus, although one cannot discard the possibility that non-conserved phosphorylations are functional, many resources use the conservation level of the site [58–60] and even of the protein [26, 61] to assess the functionality of PTMs.

The fact that phosphorylation is by far the most explored modification type also contributes to the special attention that it receives; in the future it is expected that other PTMs will be recognized as important sources of protein regulation [62], and so they will be also subject of this type of questioning. On the top of this, the conservation level of a modification site is valuable information in itself, as it reflects the evolutionary constraints of the function (if any) performed by the moiety.

The PTMcode database uses its own conservation algorithm, the *relative residue conservation score* (rRCS), to guide users in assessing the evolutionary constraints of a modification and its biological relevance. The rRCS of a PTM reports the conservation of the modified residue over orthologous proteins and is calculated as follow:

1. The protein in which the modification has been found is assigned to the oldest eukaryotic group of orthologous proteins provided by the eggNOG database [34].
2. Using the multiple alignment of the OG, the residue conservation score (RCS) is calculated for the residue. The RCS is the result of multiplying two components, the residue conservation ratio (RCR) that is the ratio of conserved sites and non-conserved sites, and the maximum branch length (MBL) of any two species containing the same residue as the PTM site from a species tree generated out of marker genes.
3. The modified residue is assigned to either an ordered or disordered region on the protein using DisEMBL [63].
4. The RCS is calculated for all residues in the OG of the same type of amino acid as the modified residue that are also in the same type of protein region (ordered or disordered). The set of scores generated here represents the background distribution used to calculate the rRCS.
5. The rRCS of the modified residue is calculated as the percentile of its RCS value in the background distribution. An rRCS >95 means that the modified residue is more conserved than the 95 % of the same amino acids within the same type of protein region.

For full details on rRCS algorithm and performance see [9]. Other people use rRCS for the same purpose [64, 65], and other resources have other types of conservation measurements based on their own algorithms [20] or on the visual inspection of orthologous protein alignments [2, 3].

Other types of data that could help to determine the biological relevance of a PTM, in the absence of specific annotation, are the number of publications where a particular PTM has been reported, the number of coevolving residues or the number of databases from which the PTM has been extracted. Indeed, in [1] we already

showed a significant positive correlation between the number of papers reporting a PTM, the normalized number of coevolving residues, and its conservation. This and other relevant information can be explored within the pop-up box that appears when the mouse is over the modification in the “protein panel.” Other clues provided by PTMcode that can be used for the assessment of the biological importance of a PTM include (1) whether there is annotation concerning the enzyme(s) that perform the modification, (2) whether the modification is inside a hotspot region, or (3) whether the modification is in a globular domain as these domains are ordered regions and so are under more evolutionary constraints.

3.7 Exploring the Functional Associations Between PTMs

The known and predicted functional associations between PTMs within a particular protein can be explored in detail from the “protein panel” (Fig. 4). From the top menu of the frame, users can display all the associations for all the PTMs. A set of connecting lines will be displayed to illustrate the predicted cross talks. Propagated PTMs (if shown) can participate in functional associations among themselves and with experimentally verified PTMs. When a single PTM is clicked, the associations shown will be restricted to those involving that modification (Fig. 4a). The lines connecting PTMs are color coded according to their association type (*see* the “evidence type” menu for details (Fig. 4b)). Clicking an “evidence” square will display only the associations with the selected evidence type; a second click deactivates the selection.

3.7.1 Exploring Cross-Talk Events by Evidence Type

The evidence for cross talk between two PTMs is displayed in the “PTM associations panel.” Users may click on each of the colored boxes in the evidence column, and a pop-up box will be displayed with further information, as detailed below:

1. Coevolution channel (green box). The conservation pattern of two modified amino acids within the MSA of the OG can be explored clicking on the “display in Jalview” button (Fig. 4c) in the pop-up box. In addition, we provide a list of species where the two amino acids are conserved.
2. Structural distance channel (yellow box). The pop-up box shows the two PTMs and the atomic distance between them (Fig. 4d). From the table row, the “highlight in structure” button will open a Jmol plug-in where the protein 3D structure with the two modified residues highlighted may be explored using full Jmol features (Fig. 4e).
3. Same residue channel (pink box). The pop-up box shows the two modified amino acids and a general annotation of their cross talk based on the types of PTMs involved. For instance, sites with a phosphorylation and an O-linked glycosylation reported are annotated as “competition” and sites with an O-linked glycosylation and a hydroxylation are annotated as “cooperation” [66].

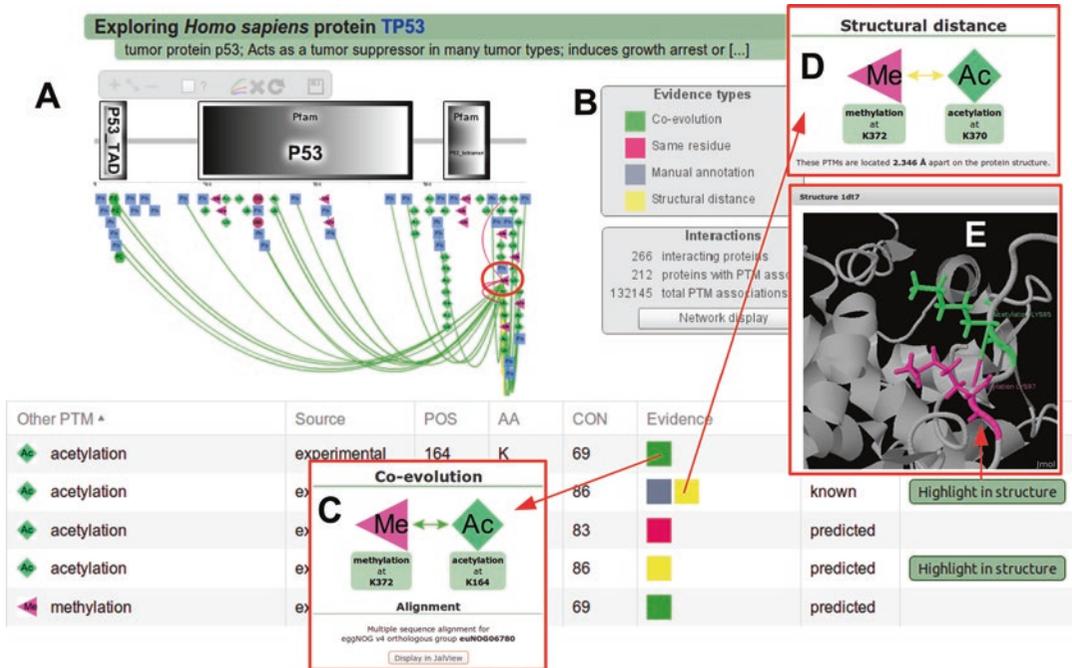


Fig. 4 Evidence channels for PTM functional associations. The human protein TP53 is methylated at position K372 (a). This methylation shows several functional associations with other PTMs supported by different “evidence channels” (b). For instance, it is found to be coevolving with the K164 residue that might be acetylated (c) and is found to be in contact with the K370 residue (e), also acetylated. This cross talk has been already reported in a scientific paper as indicated by the *gray square* (manual annotation) in the corresponding row of the table of functional associations

4. Manual annotation channel (grey box). The pop-up box shows the two modified residues, a link to the paper where the association has been reported and a single-sentence annotation that summarizes the effect of the cross talk.

3.7.2 Exploring PTM Hotspots

Inside the “protein panel,” hotspots—regions with a significant concentration of PTMs—are indicated by a blue line (Fig. 5). Clicking on the blue line will highlight the region and display the included PTMs. A click on any of the PTMs will highlight the corresponding entry in the “protein panel.” Hotspots that consist of “propagated” PTMs are only displayed if the “display propagated PTMs” checkbox is activated. To view an example of a PTM hotspot, users may check human cyclin-dependent kinase 12 (CRKRS). Interestingly, the homologous mouse protein has a hotspot in almost the same region, while in the rat protein, this region only appears as a hotspot if the “propagated” PTMs are displayed (Fig. 5). This observation highlights the need for annotation of “propagated” PTMs in species with no HTEs and supports the methodology that we use to calculate them.

CRKS orthologous proteins

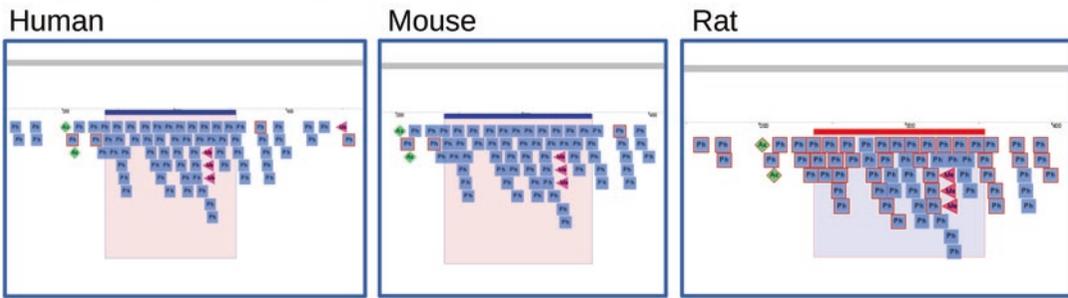


Fig. 5 PTM Hotspot representation. The hotspots found in three orthologous proteins are shown. Hotspots supported by experimentally verified PTMs are shown in *blue*; hotspots supported by “propagated” PTMs are shown in *red*

3.7.3 Export of Functional Association Data

The “protein panel” display can be downloaded using the top menu in png and jpg formats with a range of resolutions. The functional associations in text format can also be downloaded following the link below the “protein panel.”

3.8 Functional Associations of PTMs in Interacting Proteins

As discussed earlier in this Chapter, PTMs and PPIs constitute an intricate layer of protein regulation that only recently has been connected using large-scale approaches. Several computational studies have shown enrichments in PTM clusters in protein complexes [18], a higher number of interaction partners in modified proteins compared to non-modified [67] and a higher degree of coevolution between PTMs located in interacting proteins compared to non-interacting [1]. On top of these simple associations, the “PTMcode” also plays a role in the regulation of protein interactions. For example, proteins with particular coevolving PTM types form bigger protein-protein networks than proteins with same type of PTMs that are not coevolving [9].

In order to predict functional associations between PTMs in interacting proteins, some of the channels described in Subheading 3.3 were adapted to address the particularities of this task. For the “coevolution channel,” the two interacting proteins are mapped to their respective OGs, and their MSAs are pruned to keep only proteins of species in common to both. Then, the MI algorithm described in Subheading 3.3.1 is applied. For the “structural distance channel,” we measure the distance of all pairs of modified residues in protein interfaces mapped in the structure of the protein complex (if available). The “same residue channel” and “hotspot” evidence sources are not applicable here, and manual annotation was not performed for this type of association. Be aware that we apply these predictions to all the possible pairs of PTMs between the two proteins, not only those located in the binding interface, so especially for the coevolution channel, the associations may encompass a wide variety of mechanisms.

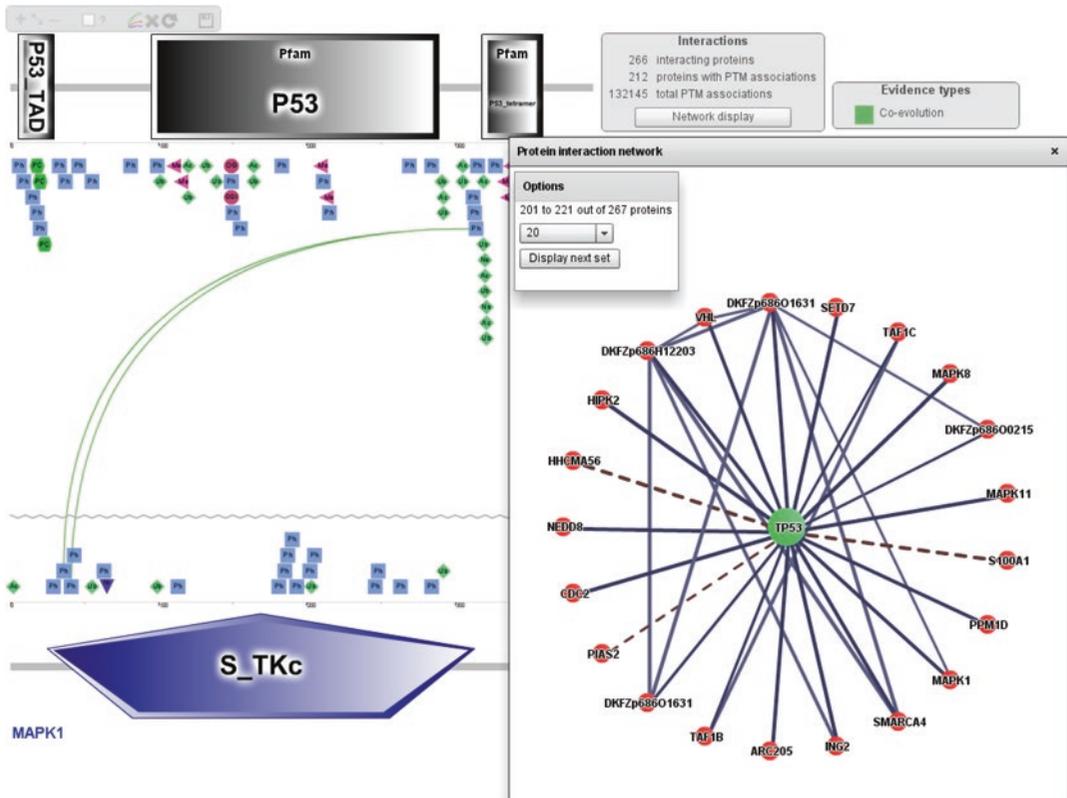


Fig. 6 Functional associations of PTMs in interacting proteins. The human protein TP53 has many reported interactions. The interaction with MAPK1 is shown within the “protein panel”; from here predicted functional associations between their PTMs on MAPK1 and TP53 can be explored in detail

The functional associations of PTMs in interacting proteins can be explored from two entry points:

1. From the home page (*see* Subheading 3.4), the tab “explore a protein pair” allows the user to display the list of interacting partners of his/her favorite protein. From that list, a particular interacting protein can be selected, and the two proteins are displayed in parallel within the “protein panel.”
2. If a single protein is being explored, as described in previous Subheadings, the “interactions menu” provides information about PPIs (*see* our definition of PPI in Subheading 3.1) and their predicted functional associations. The “network display” button will open a new panel with a representation of the PPI network (Fig. 6). From here, any PPI with a continuous edge can be explored (dashed edges represent PPIs with no PTMs functionally linked).

Once a PPI is selected, the “protein panel” shows now the two proteins (Fig. 6), their PTMs, and their functional associations. Clicking on any of them displays evidence and features as described for single proteins in previous Subheadings.

4 Notes

1. For a general review on specific cross talk between different PTM types, we suggest reading the supplementary material in [9].
2. To download the whole dataset of PTMs and functional associations, users may visit the “data” tab at the top of any PTMcode page. A direct download link is provided for associations regulating proteins and PPIs.

References

1. Minguez P, Letunic I, Parca L et al (2015) PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic Acids Res* 43:D494–D502. doi:10.1093/nar/gku1081
2. Gnad F, Gunawardena J, Mann M (2010) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* 39:D253–D260. doi:10.1093/nar/gkq1159
3. Lu C-T, Huang K-Y, Su M-G et al (2013) DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res* 41:D295–D305. doi:10.1093/nar/gks1229
4. Sadowski I, Breitzkreutz B-J, Stark C et al. (2013) The PhosphoGRID *Saccharomyces cerevisiae* protein phosphorylation site database: version 2.0 update. *Database (Oxford)* 2013:bat026. doi:10.1093/database/bat026
5. Naegle KM, Gymrek M, Joughin BA et al (2010) PTMScout, a web resource for analysis of high throughput post-translational proteomics studies. *Mol Cell Proteomics* 9:2558–2570. doi:10.1074/mcp.M110.001206
6. Craveur P, Rebehmed J, de Brevern AG (2014) PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. *Database (Oxford)* 2014:bau041. doi:10.1093/database/bau041
7. Hunter T (2007) The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol Cell* 28:730–738. doi:10.1016/j.molcel.2007.11.019
8. Beltrao P, Trinidad JC, Fiedler D et al (2009) Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol* 7:e1000134. doi:10.1371/journal.pbio.1000134
9. Minguez P, Parca L, Diella F et al (2012) Deciphering a global network of functionally associated post-translational modifications. *Mol Syst Biol* 8:599. doi:10.1038/msb.2012.31
10. The UniProt Consortium (2014) UniProt: a hub for protein information. *Nucleic Acids Res* 43:D204–D212. doi:10.1093/nar/gku989
11. Hornbeck PV, Zhang B, Murray B et al (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 43:D512–D520. doi:10.1093/nar/gku1267
12. Filtz TM, Vogel WK, Leid M (2014) Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol Sci* 35:76–85. doi:10.1016/j.tips.2013.11.005
13. Sun B, Zhang M, Cui P et al (2015) Nonsynonymous single-nucleotide variations on some posttranslational modifications of human proteins and the association with diseases. *Comput Math Methods Med* 2015:124630. doi:10.1155/2015/124630
14. Duan G, Walther D (2015) The roles of post-translational modifications in the context of protein interaction networks. *PLoS Comput Biol* 11:e1004049. doi:10.1371/journal.pcbi.1004049
15. Park CY, Krishnan A, Zhu Q et al (2014) Tissue-aware data integration approach for the inference of pathway interactions in metazoan organisms. *Bioinformatics* 31:1093–1101. doi:10.1093/bioinformatics/btu786
16. von Appen A, Kosinski J, Sparks L et al (2015) In situ structural analysis of the human nuclear pore complex. *Nature* 526:140–143. doi:10.1038/nature15381
17. Huang Y, Xu B, Zhou X et al (2015) Systematic characterization and prediction of post-translational modification cross-talk. *Mol Cell Proteomics* 14:761–770. doi:10.1074/mcp.M114.037994
18. Woodsmith J, Kamburov A, Stelzl U (2013) Dual coordination of post translational modifications in human protein networks.

- PLoS Comput Biol 9:e1002933. doi:[10.1371/journal.pcbi.1002933](https://doi.org/10.1371/journal.pcbi.1002933)
19. Creixell P, Linding R (2012) Cells, shared memory and breaking the PTM code. *Mol Syst Biol* 8:598
 20. Dinkel H, Chica C, Via A et al (2011) Phospho. ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res* 39:D261–D267. doi:[10.1093/nar/gkq1104](https://doi.org/10.1093/nar/gkq1104)
 21. Orchard S, Ammari M, Aranda B et al (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42:D358–D363. doi:[10.1093/nar/gkt1115](https://doi.org/10.1093/nar/gkt1115)
 22. Keshava Prasad TS, Goel R, Kandasamy K et al (2009) Human protein reference database—2009 update. *Nucleic Acids Res* 37:D767–D772. doi:[10.1093/nar/gkn892](https://doi.org/10.1093/nar/gkn892)
 23. Szklarczyk D, Franceschini A, Wyder S et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43:D447–D452. doi:[10.1093/nar/gku1003](https://doi.org/10.1093/nar/gku1003)
 24. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42:D191–D198. doi:[10.1093/nar/gkt1140](https://doi.org/10.1093/nar/gkt1140)
 25. Danielsen JMR, Sylvestersen KB, Bekker-Jensen S et al (2011) Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. *Mol Cell Proteomics* 10:M110.003590. doi:[10.1074/mcp.M110.003590](https://doi.org/10.1074/mcp.M110.003590)
 26. Choudhary C, Kumar C, Gnad F et al (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* 325:834–840. doi:[10.1126/science.1175371](https://doi.org/10.1126/science.1175371)
 27. Henriksen P, Wagner SA, Weinert BT et al (2012) Proteome-wide analysis of lysine acetylation suggests its broad regulatory scope in *Saccharomyces cerevisiae*. *Mol Cell Proteomics* 11:1510–1522. doi:[10.1074/mcp.M112.017251](https://doi.org/10.1074/mcp.M112.017251)
 28. Lundby A, Secher A, Lage K et al (2012) Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nat Commun* 3:876. doi:[10.1038/ncomms1871](https://doi.org/10.1038/ncomms1871)
 29. Matic I, Schimmel J, Hendriks IA et al (2010) Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. *Mol Cell* 39:641–652. doi:[10.1016/j.molcel.2010.07.026](https://doi.org/10.1016/j.molcel.2010.07.026)
 30. Murray CI, Kane LA, Uhrigshardt H et al (2011) Site-mapping of in vitro S-nitrosation in cardiac mitochondria: implications for cardioprotection. *Mol Cell Proteomics* 10:M110.004721. doi:[10.1074/mcp.M110.004721](https://doi.org/10.1074/mcp.M110.004721)
 31. Weinert BT, Wagner SA, Horn H et al (2011) Proteome-wide mapping of the Drosophila acetylome demonstrates a high degree of conservation of lysine acetylation. *Sci Signal* 4:ra48. doi:[10.1126/scisignal.2001902](https://doi.org/10.1126/scisignal.2001902)
 32. Zielinska DF, Gnad F, Schropp K et al (2012) Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. *Mol Cell* 46:542–548. doi:[10.1016/j.molcel.2012.04.031](https://doi.org/10.1016/j.molcel.2012.04.031)
 33. Wagner SA, Beli P, Weinert BT et al (2011) A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol Cell Proteomics* 10:M111.013284. doi:[10.1074/mcp.M111.013284](https://doi.org/10.1074/mcp.M111.013284)
 34. Powell S, Forslund K, Szklarczyk D et al (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* 42:D231–D239. doi:[10.1093/nar/gkt1253](https://doi.org/10.1093/nar/gkt1253)
 35. Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41:D808–D815. doi:[10.1093/nar/gks1094](https://doi.org/10.1093/nar/gks1094)
 36. Boekhorst J, van Breukelen B, Heck A, Snel B (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol* 9:R144. doi:[10.1186/gb-2008-9-10-r144](https://doi.org/10.1186/gb-2008-9-10-r144)
 37. Chen SC-C, Chen F-C, Li W-H (2010) Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals. *Mol Biol Evol* 27:2548–2554. doi:[10.1093/molbev/msq142](https://doi.org/10.1093/molbev/msq142)
 38. Tan CSH, Bodenmiller B, Pasculescu A et al (2009) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal* 2:ra39. doi:[10.1126/scisignal.2000316](https://doi.org/10.1126/scisignal.2000316)
 39. Humphrey SJ, James DE, Mann M (2015) Protein phosphorylation: a major switch mechanism for metabolic regulation. *Trends Endocrinol Metab* 26:676–687. doi:[10.1016/j.tem.2015.09.013](https://doi.org/10.1016/j.tem.2015.09.013)
 40. Byeon I-JL, Li H, Song H et al (2005) Sequential phosphorylation and multisite interactions characterize specific target recognition by the FHA domain of Ki67. *Nat Struct Mol Biol* 12:987–993. doi:[10.1038/nsmb1008](https://doi.org/10.1038/nsmb1008)
 41. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. *Nat Rev Genet* 14:249–261. doi:[10.1038/nrg3414](https://doi.org/10.1038/nrg3414)

42. Cover TM, Thomas JA (1991) Elements of information theory. John Wiley & Sons, New York
43. Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21:4116–4124. doi:[10.1093/bioinformatics/bti671](https://doi.org/10.1093/bioinformatics/bti671)
44. Skerker JM, Perchuk BS, Siryaporn A et al (2008) Rewiring the specificity of two-component signal transduction systems. *Cell* 133:1043–1054. doi:[10.1016/j.cell.2008.04.040](https://doi.org/10.1016/j.cell.2008.04.040)
45. Pietal MJ, Bujnicki JM, Kozlowski LP (2015) GDFuzz3D: a method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function. *Bioinformatics* 31:3499–3505. doi:[10.1093/bioinformatics/btv390](https://doi.org/10.1093/bioinformatics/btv390)
46. Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The future of the protein data bank. *Polymers*. doi:[10.1002/bip.22132](https://doi.org/10.1002/bip.22132)
47. Seet BT, Dikic I, Zhou M-M, Pawson T (2006) Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* 7:473–483. doi:[10.1038/nrml960](https://doi.org/10.1038/nrml960)
48. Hart GW, Greis KD, Dong LY et al (1995) O-linked N-acetylglucosamine: the “yin-yang” of Ser/Thr phosphorylation? Nuclear and cytoplasmic glycosylation. *Adv Exp Med Biol* 376:115–123
49. Latham JA, Dent SYR (2007) Cross-regulation of histone modifications. *Nat Struct Mol Biol* 14:1017–1024. doi:[10.1038/nsmb1307](https://doi.org/10.1038/nsmb1307)
50. Brooks CL, Gu W (2003) Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr Opin Cell Biol* 15:164–171. doi:[10.1016/S0955-0674\(03\)00003-6](https://doi.org/10.1016/S0955-0674(03)00003-6)
51. Beltrao P, Albanèse V, Kenner LR et al (2012) Systematic functional prioritization of protein posttranslational modifications. *Cell* 150:413–425. doi:[10.1016/j.cell.2012.05.036](https://doi.org/10.1016/j.cell.2012.05.036)
52. Zeidan Q, Hart GW (2010) The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. *J Cell Sci* 123:13–22. doi:[10.1242/jcs.053678](https://doi.org/10.1242/jcs.053678)
53. Butt AM, Khan IB, Hussain M et al (2011) Role of post translational modifications and novel crosstalk between phosphorylation and O-beta-GlcNAc modifications in human claudin-1, -3 and -4. *Mol Biol Rep* 39:1359–1369. doi:[10.1007/s11033-011-0870-7](https://doi.org/10.1007/s11033-011-0870-7)
54. Vodermaier HC (2004) APC/C and SCF: controlling each other and the cell cycle. *Curr Biol* 14:R787–R796. doi:[10.1016/j.cub.2004.09.020](https://doi.org/10.1016/j.cub.2004.09.020)
55. Letunic I, Doerks T, Bork P (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* 43:D257–D260. doi:[10.1093/nar/gku949](https://doi.org/10.1093/nar/gku949)
56. Lienhard GE (2008) Non-functional phosphorylations? *Trends Biochem Sci* 33:351–352. doi:[10.1016/j.tibs.2008.05.004](https://doi.org/10.1016/j.tibs.2008.05.004)
57. Wang Z, Ding G, Geistlinger L et al (2011) Evolution of protein phosphorylation for distinct functional modules in vertebrate genomes. *Mol Biol Evol* 28:1131–1140. doi:[10.1093/molbev/msq268](https://doi.org/10.1093/molbev/msq268)
58. Gnad F, Ren S, Cox J et al (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8:R250. doi:[10.1186/gb-2007-8-11-r250](https://doi.org/10.1186/gb-2007-8-11-r250)
59. Holt LJ, Tuch BB, Villén J et al (2009) Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325:1682–1686. doi:[10.1126/science.1172867](https://doi.org/10.1126/science.1172867)
60. Tan CSH, Bader GD (2012) Phosphorylation sites of higher stoichiometry are more conserved. *Nat Methods* 9:317. doi:[10.1038/nmeth.1941](https://doi.org/10.1038/nmeth.1941)
61. Zielinska DF, Gnad F, Wiśniewski JR, Mann M (2010) Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* 141:897–907. doi:[10.1016/j.cell.2010.04.012](https://doi.org/10.1016/j.cell.2010.04.012)
62. Martínez-Ruiz A, Lamas S (2004) S-nitrosylation: a potential new paradigm in signal transduction. *Cardiovasc Res* 62:43–52. doi:[10.1016/j.cardiores.2004.01.013](https://doi.org/10.1016/j.cardiores.2004.01.013)
63. Linding R, Jensen LJ, Diella F et al (2003) Protein disorder prediction: implications for structural proteomics. *Structure* 11:1453–1459
64. Ullah S, Lin S, Xu Y et al (2016) dbPAF: an integrative database of protein phosphorylation in animals and fungi. *Sci Rep* 6:23534. doi:[10.1038/srep23534](https://doi.org/10.1038/srep23534)
65. Pan Z, Liu Z, Cheng H et al (2014) Systematic analysis of the in situ crosstalk of tyrosine modifications reveals no additional natural selection on multiply modified residues. *Sci Rep* 4:7331. doi:[10.1038/srep07331](https://doi.org/10.1038/srep07331)
66. Wang ZA, Singh D, van der Wel H, West CM (2011) Prolyl hydroxylation- and glycosylation-dependent functions of Skp1 in O₂-regulated development of *Dictyostelium*. *Dev Biol* 349:283–295. doi:[10.1016/j.ydbio.2010.10.013](https://doi.org/10.1016/j.ydbio.2010.10.013)
67. Yachie N, Saito R, Sugiyama N et al (2011) Integrative features of the yeast phosphoproteome and protein–protein interaction map. *PLoS Comput Biol* 7:e1001064. doi:[10.1371/journal.pcbi.1001064](https://doi.org/10.1371/journal.pcbi.1001064)

Bioinformatics Analysis of PTM-Modified Protein Interaction Networks and Complexes

Jonathan Woodsmith, Ulrich Stelzl, and Arunachalam Vinayagam

Abstract

Normal cellular functioning is maintained by macromolecular machines that control both core and specialized molecular tasks. These machines are in large part multi-subunit protein complexes that undergo regulation at multiple levels, from expression of requisite components to a vast array of post-translational modifications (PTMs). PTMs such as phosphorylation, ubiquitination, and acetylation currently number more than 200,000 in the human proteome and function within all molecular pathways. Here we provide a framework for systematically studying these PTMs in the context of global protein–protein interaction networks. This analytical framework allows insight into which functions specific PTMs tend to cluster in, and furthermore which complexes either single or multiple PTM signaling pathways converge on.

Key words Protein interaction networks, Post-translational modifications, Protein complexes, Modification cross talk

1 Introduction

Most cellular functions are driven by the coordinated action of multiple macromolecular assemblies of interacting protein subunits. Defining the molecular architecture of how these individual protein building blocks interact is a major task fundamental to a better understanding of cellular processes in health and disease [1–3]. Recent protein–protein interaction (PPI) studies have improved interactome data coverage and provided novel insight into multiple cellular systems [4–8].

Generating datasets broad in scope is fundamental to interactome mapping, providing an increasingly better framework for further analysis. Much work to improve data quality had focused previously on determining and improving the specificity of large-scale PPI approaches [9–11]. However, recent methods have substantially improved issues surrounding poor interaction space coverage; e.g., yeast two-hybrid approaches have adopted second-generation

sequencing techniques to reduce workload and increase sensitivity in large-scale yeast two-hybrid dataset generation [12, 13]. The yeast two-hybrid PPI search space has been expanded toward modification-dependent protein interactions [14] and disease variant proteins [15, 16]. Mass spectrometry-based PPI approaches have become more quantitative [6] and comparative [8]. These interaction datasets, along with other recent well-controlled high-throughput studies, are invaluable for their systematic approach and have built a relatively unbiased framework for the analyses of genetic variation, transcription data, protein expression, and post-translational modification datasets [2].

As such, there have been recent efforts to curate datasets that integrate multiple sources of PPIs [4, 17–21]. One related approach, termed COMPLETEAT, developed a framework for the analysis of high-throughput datasets at the level of protein complexes [22]. Protein complexes are the functional units of proteome organization, and their dynamic assembly is fundamental in inducing rapid cellular responses to different internal and external cues. For datasets that include multiple conditions or time points, a protein complex-based analysis might be preferable because it could reveal network dynamics that are missed in other types of analyses. Moreover, the individual protein complexes that participate in a signaling pathway assemble in different compartments and at different times, and selected complexes associated with a pathway might integrate signals from other pathways. Finally protein complexes do not rely on prior knowledge or functional annotations as such and thus provide a broader more unbiased basis for functional analysis.

We suggest that protein–protein interaction networks, such as protein complex datasets, are very useful in gaining a more comprehensive understanding of post-translational modifications. Human PTMs are known to number greater than 400 and range from chemical modifications of amino acid side chains such as acetylation and phosphorylation, to the addition of the large peptide chains of the ubiquitin and ubiquitin-like families through isopeptide bonds [23]. PTM regulation is achieved by a large number of components encoded by 5–10 % of the protein coding genome, each controlled by distinct regulatory systems that vary in both number and mechanism of modification. For example, reversible protein phosphorylation is controlled through the direct action of >500 kinases and >150 phosphatases in human. The critical requirement for normal PTM functioning can be observed as many regulatory proteins are annotated in disease pathogenesis and are the targets of current drugs in ongoing clinical trials [24, 25].

Large datasets recording post-translational modifications via mass spectrometry have been made available. For example, phosphoproteomics analysis of mammalian samples typically record

more than 1000, sometimes up to 50,000 sites [26]. Several differential datasets demonstrate that hundreds of phosphosites change specifically in response to an external trigger or a cellular perturbation. Even though stoichiometry of the phosphosite may be low and the functional impact (effect size) of single modifications of proteins on the phenotype may vary substantially, it is clear that any given phospho-response of the cell is widespread and affects many cellular pathways and machines [27, 28].

Systematic analysis of PTM regulation involving multiple signaling PTMs, such as phosphorylation, acetylation, ubiquitination, and methylation, had been limited by data paucity. In particular, studies addressing more than one PTM type at a time [29] are largely elusive. However, in recent years, serine/threonine and tyrosine phosphorylation datasets have been augmented by acetylation and ubiquitination datasets to provide a more comprehensive view of cellular signaling events. These systematic datasets have allowed global analysis both of individual PTM traits and more interestingly their interconnectivity and relationships to one another within individual proteins and across both evolutionary and protein-protein interaction frameworks [30–40].

Here we detail how to integrate PTM data with a protein complex interaction framework. Part of the analysis has been carried out in the COMPLETEAT web-based tool. Building this framework not only allows for systematic analysis of novel PTM datasets but allows the contextualization of individual experimental PTM datasets.

2 Materials

2.1 Protein Interaction Data from Distinct Sources

Experimentally identified PPI for human, *Drosophila*, and yeast were systematically compiled from major PPI databases such as BioGrid [18], IntAct, and MINT [19]. Further, these resources were complemented with species-specific interaction databases such as DroID [41] for *Drosophila*. The PPIs were downloaded from the corresponding source databases and the gene/protein identifiers were mapped to Entrez gene identifiers. Next these PPI resources were integrated to build a comprehensive PPI network for human, *Drosophila*, and yeast [22]. As a complement to the curated resources, we also used a human interaction dataset retrieved from a collection of more than 50 high-quality, large-scale interaction studies [4] containing more than 100,000 human PPIs which should be less affected by research biases.

2.2 Post-translational Modification Datasets

Data for each PTM was obtained from PhosphositePlus [42] and integrated with publically available datasets (available as supplementary tables from specific papers) to obtain a nonredundant list

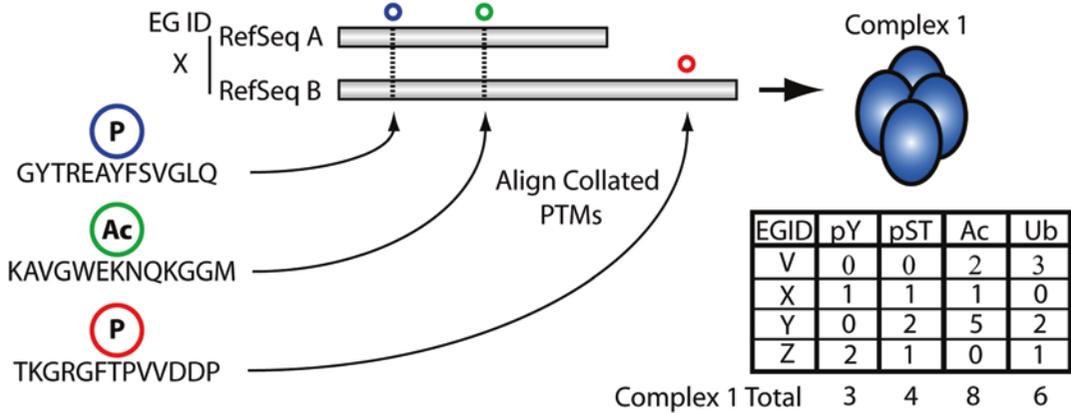


Fig. 1 Mapping post-translational modifications to unique sequence identifiers and proteins in protein complexes. V, X, Y, and Z are representative of Entrez gene identifiers that in turn point to the specific protein products that make up any given complex

of modified 13 amino acid sequences (13mers). The central amino acid was annotated as modified in each 13mer and only modified tyrosine, lysine, serine, or threonine residues were utilized in the final analysis. The numerical position of each annotated PTM was critical for downstream analysis as integrating multiple datasets can cause errors in modified residue placement. This can be either due to different isoform/protein sequences being utilized in different studies, which would lead to differing numbering systems, or simple annotation errors. Most studies provide protein sequence information surrounding the annotated site that can be used to control for position within the specific isoform used. Therefore, each 13mer is standardized in annotation type through sequence-specific matching to an individual RefSeq protein sequence, and the most highly annotated RefSeq protein is used for any given Entrez gene identifier (Fig. 1). In order to be able to combine the PTM dataset with the interaction dataset, when multiple very similar protein products are annotated across many RefSeq sequences and Entrez gene identifiers, the most highly annotated protein in each class was selected for further analysis as representative of the protein subfamily (*see Note 1*). These careful control steps are taken to allow robust analysis of PTMs within a single protein and across many protein complexes, reducing annotation and redundancy errors that can occur when collating multiple datasets (Fig. 1). An example dataset of 100,391 uniquely mapped 13mers annotated across 12,127 unique proteins is freely available online (Dataset S1 in [32]) (*see Note 2*).

3 Methods

3.1 Building a Network of Complexes from Varied Interaction Data Sources

To build a comprehensive protein complex resource for human, *Drosophila*, and *S. cerevisiae*, two different approaches were combined (Fig. 2). First, protein complexes from literature-curated resources (referred as literature-based protein complexes) were compiled. This literature-based protein complexes resource was then supplemented with high-confidence putative protein complexes identified from protein–protein interaction networks. Combined, the literature-based and deduced protein complexes create the largest repository available of complexes for human, *Drosophila*, or yeast.

Binary interaction information was obtained from resources described in Subheading 2.1 (PPI networks in Fig. 2). Literature-based complexes were compiled from databases such as CORUM, PINdb, CYC2008, gene ontology (GO), KEGG, and DPiM. With the exception of protein complexes that are annotated by GO, all the other complexes are mapped across human, *Drosophila*, and yeast using DIOPT (<http://www.flyrnai.org/diopt>), an integrative ortholog prediction tool. DIOPT scores were used to select the best ortholog match in case of “one-to-many” ortholog relationships. Only complexes consisting of two or more proteins were included in the complex resources generated. Complex annotations from the source databases, including complex name, purification method, and PubMed ID, were also included.

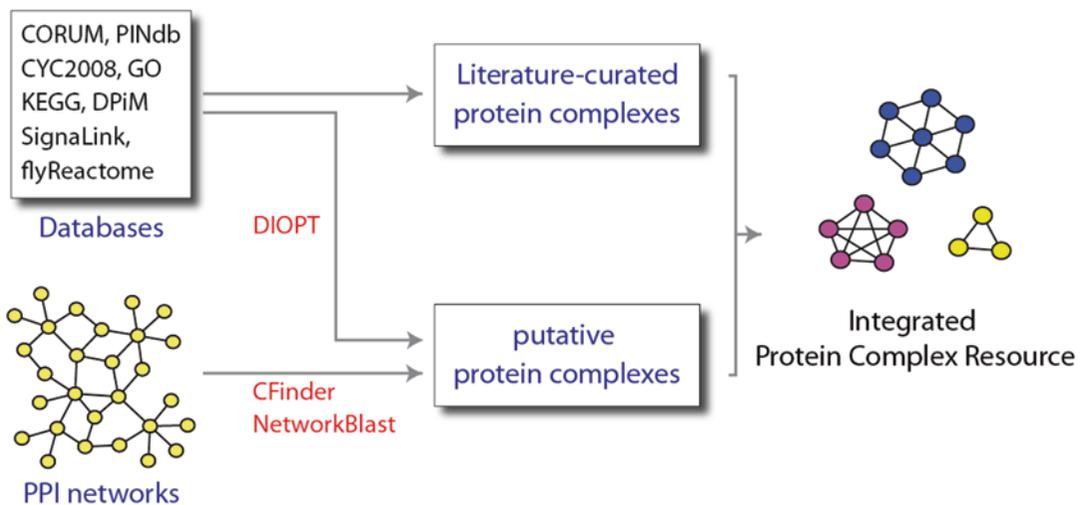


Fig. 2 Building comprehensive protein complex resources available via COMPLEAT at <http://www.flyrnai.org/compleat/>

3.2 Predicted Protein Complexes

Two complementary computational tools were used to predict protein complexes from PPI networks: CFinder [43] and NetworkBlast [44]. CFinder identifies protein complexes from a single PPI network, whereas NetworkBlast is a network alignment tool that identifies conserved protein complexes by aligning two or more networks. CFinder (<http://www.cfinder.org/>) can be downloaded and implemented locally. CFinder uses the clique percolation method (CPM) to locate k-clique communities. A k-clique community is a union of all smaller complete subgraphs (or k-cliques) that can be reached from each other through a series of adjacent k-cliques. We used unweighted, undirected PPI networks as input from human, *Drosophila*, and yeast. We filtered the PPI networks using co-expression values or co-localization information to remove low-confidence PPIs. Co-expression values were used to filter human and *Drosophila* PPI networks, and only edges with Pearson correlation ≥ 0.2 were retained. For the yeast network, we used co-localization information and retained only those PPIs where the subcellular localization of both proteins is known and the proteins co-localize. We used the filtered human, *Drosophila*, and yeast PPI networks as input and ran CFinder with the default parameters. NetworkBLAST was used to identify evolutionarily conserved protein complexes by aligning two networks from different species. NetworkBLAST can be downloaded (<http://www.cs.tau.ac.il/~bnet/networkblast.htm>) and implemented locally. We used stringent parameters to align the human, *Drosophila*, and yeast networks. The complex density was set to 0.95 and false negatives to 0.2, 0.2, and 0.1 for the human, *Drosophila*, and yeast networks respectively.

3.3 Studying Enrichment of PTMs Within a Network of Protein Complexes

PTMs are mapped via proteins to the human COMPLEAT complex dataset for enrichment analysis. Each gene product in the protein complex and PTM datasets is annotated with a unique Entrez gene identifier. It is critical that each unique gene ID points to a unique protein sequence (here RefSeq identifiers, Fig. 1). Currently, we ignore isoform level annotation, because this information across multiple data types is very sparse (*see Note 3*). Furthermore, it allows nonredundant, rapid, and error-free annotation of PTMs across each protein complex using well-defined lookup tables (Fig. 3).

Protein complex size varies considerably, both in number of unique protein subunits and total protein content. It was therefore necessary to assess the enrichment of PTMs within the network over a null PTM distribution randomized model. Random protein complex dataset generation was undertaken via PTM annotation permutation performed at the protein level, keeping all modifications across an identifier linked when shuffled. Annotation permutation was performed within 16 individual bins of approximately equal size; each bin contained proteins within the same size and complex dataset frequency quartiles to account for slight correlations observed in the dataset characterization. One hundred annotation

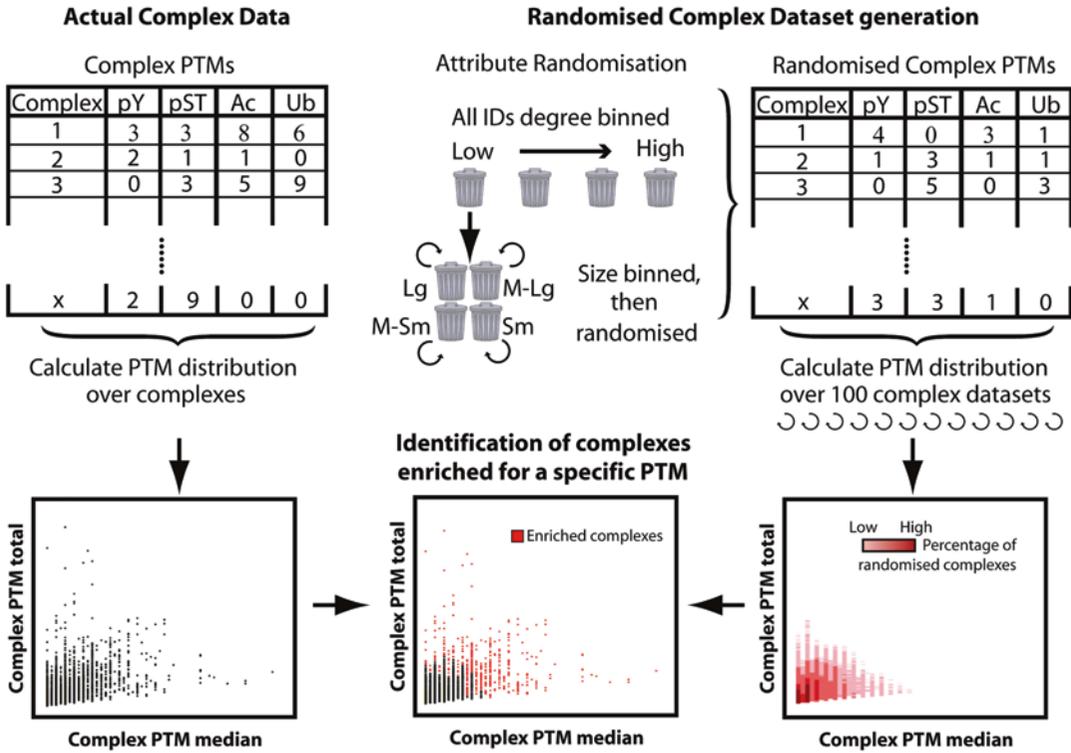


Fig. 3 Identifying highly modified protein complexes in comparison to protein complex data with randomized protein annotation. In the middle panel, 1 % outliers of the distribution are color coded in *red*. For any given modification, these are typically hundreds of human complexes that are highly modified and can support hypothesis generation

randomized datasets were collated and the data distribution for each modification ascertained, giving an indication of what modification levels would be across complexes by random chance. To visualize highly modified complexes, the total modification across a complex was plotted against the median modification per subunit and overlaid on the randomized dataset distribution (Fig. 3). As such, 1 % outliers of the distribution are identified as highly modified complexes and can either provide a basis for specific hypothesis driven experimentation on selected complexes or guide systematic functional prioritization, e.g., through GO enrichment analysis.

3.4 GO Analysis of Enriched Complexes

GO functional analysis was undertaken using the ConsensusPathPD overrepresentation analysis tool [17] using gene ontology level four categories for both biological processes and molecular functions (p -value cutoff, 0.05). Only protein identifiers present in complexes that were enriched for the designated PTMs alone were taken forward to the GO analysis, using a background dataset consisting of all identifiers found in the protein complex interaction dataset. Using this simple approach, we obtained a first look at the pathways and processes specifically enriched in a group of protein

complexes. In particular this analysis can highlight overarching functional distinctions between datasets, either between different PTMs [32], or it can be extended to comparative analysis of different stimuli, time points, or cell types. In general, p -values and distinct input lists at varying enrichment cutoffs can be used to filter results and observe GO enrichment. However, this relatively simplistic approach to assigning confidence does not always help to identify the most biologically relevant and novel findings from typically very long lists of enriched GO terms.

3.5 COMPLEAT Analysis of PTMs to Identify Complexes

Though GO and pathway annotations are extremely useful, there is a need to broaden the data analyses from a different perspective. For example, recent PTM datasets are being generated under different conditions and/or time points with the objective of capturing the dynamics of biological systems [27, 28, 45]. As a result, common challenges in analyzing PTM datasets are to distinguish signal from noise, place results in a functional context, and prioritize a subset of candidates for further validation. Several enrichment analysis methods, including GO, pathway, and gene set-based enrichment analyses, have been developed to help address these needs. For PTM data, protein complex-based analysis is more appropriate since protein complexes are the functional units of proteome organization and their dynamic assembly is fundamental in bringing about differences in cellular responses to different internal and external cues. Protein complex-based analysis could also address another limitation of the existing gene set annotations, which tend to be either too specific or too broad. In order to understand how PTMs responded at a systems level, we must be able to visualize and study the dynamics of protein complexes and cellular networks. COMPLEAT analysis was developed to address these issues.

COMPLEAT analysis is a four-step procedure that can be performed via a web-based interface at <http://www.flyrnai.org/compleat>. (1) Upload your data from a tab-delimited input file, (2) choose the organism, (3) press submit and run the program, and (4) obtain the output as graphical representation and as table view using the menu. P -value and score can be adjusted to obtain results at user-defined cutoff and redundant complexes can be masked by enabling “hide redundant.” Example files, a YouTube demo clip and a FAQ section, are available at the site.

A gene or protein list with mapped PTMs can be used as input. The tool can handle multiple identifiers as input including Entrez gene ID, gene symbol, and Uniprot ID. In addition to the gene identifier, the tool also requires a weight for each gene/protein. This could be a discrete value (“1” if there is PTM observed for given proteins or “0” if no PTM) or continuous values (e.g., number of PTMs observed in a protein or a confidence score pertaining to the observed PTMs). To find enriched complexes, i.e., complexes that

have protein members overrepresented in the PTM-input data, COMPLEAT maps the input values to each complex using gene or protein identifier and computes a score and p -value. The score corresponds to the interquartile mean (IQM) of the input values for all members mapped to the complex. To compute a p -value of a given complex size n , COMPLEAT generates 1000 randomized protein complexes with size n and computes the random distribution of IQMs. This is modeled as probability density distribution and used to compute the p -value of the given complex [22]. The results are visualized as an interactive scatter plot. Using the interactive scatter plot, users can select complexes of interest (based on complex score or p -value), and network illustrations of the selected complexes are displayed using Cytoscape Web.

In addition to analyzing a single PTM dataset, users can compare multiple PTM datasets generated under different conditions or time points. Such comparative analysis could help to understand how PTMs responded at a systems level and we must be able to visualize and study the dynamics of protein complexes and cellular networks.

3.6 Delving Deeper into the PTM Data: Distribution of PTMs Across Individual Protein Subunits

Having observed enrichment of the PTM signal across the protein complexes, we can then see whether signal enrichment can be observed at a more detailed level of individual complex subunits. Recent approaches have used protein structures to observe clustering of PTMs within three-dimensional space; however, these are necessarily restricted to the limited 3D structure models available. To study all proteins in the dataset, we undertook a simpler approach of scanning across the linear protein sequence in windows of 20AAs to ascertain whether PTM enrichment could be observed from this more detailed viewpoint (Fig. 4). To obtain local maxima of PTM density, 20AA windows are calculated with 10AA overlap to be able to ascertain the maximum PTM density over any given sequence space. PTM peaks can be further characterized in detail through comparison with protein sequence-based annotations such as amino acid content, disorder prediction (e.g., using Iupred disorder prediction software, <http://iupred.enzim.hu/>), and protein domains (e.g., from interpro, <https://www.ebi.ac.uk/interpro/>). Using this approach regions of particularly high PTM density, termed PTMi spots [32], were identified in hundreds of human proteins. PTMi spots represent very dense modification patterns in disordered protein regions and may be important for PTM crosstalk. PTMi spots showed an equally high mutation rate as functional protein domains in cancer, inferring equivocal importance for cellular functioning [32]. Understanding PTM enrichment in the context of surrounding protein features provides a greater depth of understanding and can lead directly to hypothesis driven experimentation.

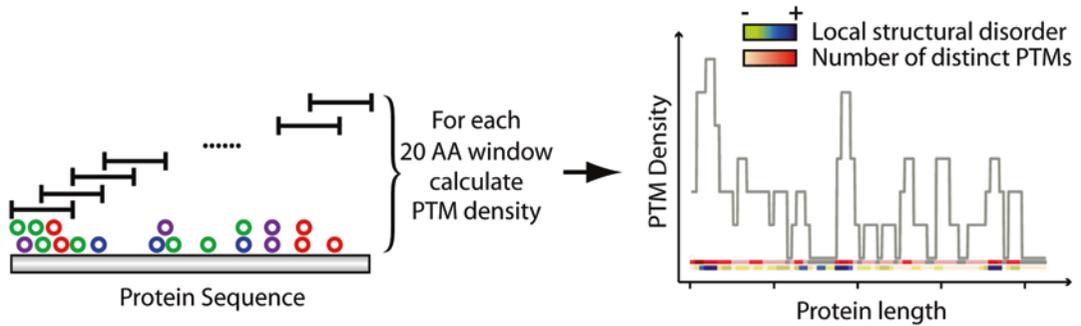


Fig. 4 Identifying PTMi spots through scanning across the linear protein sequence in windows of 20AAs

4 Notes

1. Only one RefSeq identifier per Entrez gene identifier is taken further, as protein interaction networks typically refer to one protein per GeneID. In special cases, if the modifications map to several Entrez gene identifiers, they are collapsed. For example, histones are represented by multiple unique Entrez gene identifiers, so to be able to assign PTMs to a specific Entrez gene identifier, the most highly annotated GeneID and associated RefSeq are taken forward to the final table.
2. Please be aware that post-translational modification data grow at an enormous rate. Therefore, datasets have to be assembled frequently for up to date analyses.
3. This protocol tries to circumvent the problem of different splice isoforms in the analyses by using the form that is most highly modified by PTMs. However, even though interaction networks are currently not isoform specific, they typically refer to a reference protein isoform via an Entrez gene identifier. It is also important to note that protein interactions will vary substantially between isoforms of the same protein [46].

References

1. Vidal M, Cusick ME, Barabási A (2011) Interactome networks and human disease. *Cell* 144(6):986–998. doi:[10.1016/j.cell.2011.02.016](https://doi.org/10.1016/j.cell.2011.02.016)
2. Stelzl U (2013) Molecular interaction networks in the analyses of sequence variation and proteomics data. *Proteomics Clin Appl* 7(11–12):727–732. doi:[10.1002/prca.201300039](https://doi.org/10.1002/prca.201300039)
3. Snider J, Kotlyar M, Saraon P et al (2015) Fundamentals of protein interaction network mapping. *Mol Syst Biol* 11(12):848. doi:[10.15252/msb.20156351](https://doi.org/10.15252/msb.20156351)
4. Woodsmith J, Stelzl U (2014) Studying post-translational modifications with protein interaction networks. *Curr Opin Struct Biol* 24:34–44. doi:[10.1016/j.sbi.2013.11.009](https://doi.org/10.1016/j.sbi.2013.11.009)
5. Rolland T, Taşan M, Charloteaux B et al (2014) A proteome-scale map of the human interactome network. *Cell* 159(5):1212–1226. doi:[10.1016/j.cell.2014.10.050](https://doi.org/10.1016/j.cell.2014.10.050)
6. Hein MY, Hubner NC, Poser I et al (2015) A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163(3):712–723. doi:[10.1016/j.cell.2015.09.053](https://doi.org/10.1016/j.cell.2015.09.053)

7. Huttlin EL, Ting L, Bruckner RJ et al (2015) The BioPlex network: a systematic exploration of the human interactome. *Cell* 162(2):425–440. doi:[10.1016/j.cell.2015.06.043](https://doi.org/10.1016/j.cell.2015.06.043)
8. Wan C, Borgeson B, Phanse S et al (2015) Panorama of ancient metazoan macromolecular complexes. *Nature* 525(7569):339–344. doi:[10.1038/nature14877](https://doi.org/10.1038/nature14877)
9. Venkatesan K, Rual J, Vazquez A et al (2009) An empirical framework for binary interactome mapping. *Nat Methods* 6(1):83–90. doi:[10.1038/nmeth.1280](https://doi.org/10.1038/nmeth.1280)
10. Mellacheruvu D, Wright Z, Couzens AL et al (2013) The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nat Methods* 10(8):730–736. doi:[10.1038/nmeth.2557](https://doi.org/10.1038/nmeth.2557)
11. Varjosalo M, Sacco R, Stukalov A et al (2013) Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat Methods* 10(4):307–314. doi:[10.1038/nmeth.2400](https://doi.org/10.1038/nmeth.2400)
12. Weimann M, Grossmann A, Woodsmith J et al (2013) A Y2H-seq approach defines the human protein methyltransferase interactome. *Nat Methods* 10(4):339–342. doi:[10.1038/nmeth.2397](https://doi.org/10.1038/nmeth.2397)
13. Yu H, Tardivo L, Tam S et al (2011) Next-generation sequencing to generate interactome datasets. *Nat Methods* 8(6):478–480. doi:[10.1038/nmeth.1597](https://doi.org/10.1038/nmeth.1597)
14. Grossmann A, Benlasfer N, Birth P et al (2015) Phospho-tyrosine dependent protein–protein interaction network. *Mol Syst Biol* 11(3):794. doi:[10.15252/msb.20145968](https://doi.org/10.15252/msb.20145968)
15. Sahni N, Yi S, Taipale M et al (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161(3):647–660. doi:[10.1016/j.cell.2015.04.013](https://doi.org/10.1016/j.cell.2015.04.013)
16. Wei X, Das J, Fragoza R et al (2014) A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet* 10(12):e1004819. doi:[10.1371/journal.pgen.1004819](https://doi.org/10.1371/journal.pgen.1004819)
17. Kamburov A, Stelzl U, Lehrach H et al (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 41(Database issue):D793–D800. doi:[10.1093/nar/gks1055](https://doi.org/10.1093/nar/gks1055)
18. Chatr-Aryamontri A, Breitkreutz B, Oughtred R et al (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43(Database issue):D470–D478. doi:[10.1093/nar/gku1204](https://doi.org/10.1093/nar/gku1204)
19. Orchard S, Ammari M, Aranda B et al (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42(Database issue):D358–D363. doi:[10.1093/nar/gkt1115](https://doi.org/10.1093/nar/gkt1115)
20. Schaefer MH, Fontaine J, Vinayagam A et al (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One* 7(2):e31826. doi:[10.1371/journal.pone.0031826](https://doi.org/10.1371/journal.pone.0031826)
21. Das J, Yu H (2012) HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* 6:92. doi:[10.1186/1752-0509-6-92](https://doi.org/10.1186/1752-0509-6-92)
22. Vinayagam A, Hu Y, Kulkarni M et al (2013) Protein complex-based analysis framework for high-throughput data sets. *Sci Signal* 6(264):rs5. doi:[10.1126/scisignal.2003629](https://doi.org/10.1126/scisignal.2003629)
23. Khoury GA, Baliban RC, Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep* 1. doi:[10.1038/srep00090](https://doi.org/10.1038/srep00090)
24. Cohen P, Tcherpakov M (2010) Will the ubiquitin system furnish as many drug targets as protein kinases? *Cell* 143(5):686–693. doi:[10.1016/j.cell.2010.11.016](https://doi.org/10.1016/j.cell.2010.11.016)
25. Zhang J, Yang PL, Gray NS (2009) Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer* 9(1):28–39. doi:[10.1038/nrc2559](https://doi.org/10.1038/nrc2559)
26. von L S, Francavilla C, Olsen JV (2015) Recent findings and technological advances in phosphoproteomics for cells and tissues. *Expert Rev Proteomics* 12(5):469–487. doi:[10.1586/14789450.2015.1078730](https://doi.org/10.1586/14789450.2015.1078730)
27. Olsen JV, Vermeulen M, Santamaria A et al (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal* 3(104):ra3. doi:[10.1126/scisignal.2000475](https://doi.org/10.1126/scisignal.2000475)
28. Rigbolt KT, Prokhorova TA, Akimov V et al (2011) System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Sci Signal* 4(164):rs3. doi:[10.1126/scisignal.2001570](https://doi.org/10.1126/scisignal.2001570)
29. Swaney DL, Beltrao P, Starita L et al (2013) Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. *Nat Methods* 10(7):676–682. doi:[10.1038/nmeth.2519](https://doi.org/10.1038/nmeth.2519)
30. Beltrao P, Albanèse V, Kenner LR et al (2012) Systematic functional prioritization of protein posttranslational modifications. *Cell* 150(2):413–425. doi:[10.1016/j.cell.2012.05.036](https://doi.org/10.1016/j.cell.2012.05.036)
31. Minguéz P, Parca L, Diella F et al (2012) Deciphering a global network of functionally associated post-translational modifications. *Mol Syst Biol* 8:599. doi:[10.1038/msb.2012.31](https://doi.org/10.1038/msb.2012.31)

32. Woodsmith J, Kamburov A, Stelzl U (2013) Dual coordination of post translational modifications in human protein networks. *PLoS Comput Biol* 9(3):e1002933. doi:[10.1371/journal.pcbi.1002933](https://doi.org/10.1371/journal.pcbi.1002933)
33. Lothrop AP, Torres MP, Fuchs SM (2013) Deciphering post-translational modification codes. *FEBS Lett* 587(8):1247–1257. doi:[10.1016/j.febslet.2013.01.047](https://doi.org/10.1016/j.febslet.2013.01.047)
34. Peng M, Scholten A, Heck AJR et al (2014) Identification of enriched PTM crosstalk motifs from large-scale experimental data sets. *J Proteome Res* 13(1):249–259. doi:[10.1021/pr4005579](https://doi.org/10.1021/pr4005579)
35. Winter DL, Erce MA, Wilkins MR (2014) A web of possibilities: network-based discovery of protein interaction codes. *J Proteome Res* 13(12):5333–5338. doi:[10.1021/pr500585p](https://doi.org/10.1021/pr500585p)
36. Pejaver V, Hsu W, Xin F et al (2014) The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci* 23(8):1077–1093. doi:[10.1002/pro.2494](https://doi.org/10.1002/pro.2494)
37. Yao H, Li A, Wang M (2015) Systematic analysis and prediction of in situ cross talk of O-GlcNAcylation and phosphorylation. *Biomed Res Int* 2015:279823. doi:[10.1155/2015/279823](https://doi.org/10.1155/2015/279823)
38. Dewhurst HM, Choudhury S, Torres MP (2015) Structural analysis of PTM hotspots (SAPHire)—a quantitative informatics method enabling the discovery of novel regulatory elements in protein families. *Mol Cell Proteomics* 14(8):2285–2297. doi:[10.1074/mcp.M115.051177](https://doi.org/10.1074/mcp.M115.051177)
39. Huang Y, Xu B, Zhou X et al (2015) Systematic characterization and prediction of post-translational modification cross-talk. *Mol Cell Proteomics* 14(3):761–770. doi:[10.1074/mcp.M114.037994](https://doi.org/10.1074/mcp.M114.037994)
40. Duan G, Walther D (2015) The roles of post-translational modifications in the context of protein interaction networks. *PLoS Comput Biol* 11(2):e1004049. doi:[10.1371/journal.pcbi.1004049](https://doi.org/10.1371/journal.pcbi.1004049)
41. Murali T, Pacifico S, Yu J et al (2011) DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res* 39(Database issue):D736–D743. doi:[10.1093/nar/gkq1092](https://doi.org/10.1093/nar/gkq1092)
42. Hornbeck PV, Kornhauser JM, Tkachev S et al (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40(Database issue):D261–D270. doi:[10.1093/nar/gkr1122](https://doi.org/10.1093/nar/gkr1122)
43. Adamcsek B, Palla G, Farkas IJ et al (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22(8):1021–1023. doi:[10.1093/bioinformatics/btl039](https://doi.org/10.1093/bioinformatics/btl039)
44. Kalaev M, Smoot M, Ideker T et al (2008) NetworkBLAST: comparative analysis of protein networks. *Bioinformatics* 24(4):594–596. doi:[10.1093/bioinformatics/btm630](https://doi.org/10.1093/bioinformatics/btm630)
45. Sopko R, Foos M, Vinayagam A et al (2014) Combining genetic perturbations and proteomics to examine kinase-phosphatase networks in *Drosophila* embryos. *Dev Cell* 31(1):114–127. doi:[10.1016/j.devcel.2014.07.027](https://doi.org/10.1016/j.devcel.2014.07.027)
46. Yang X, Coulombe-Huntington J, Kang S et al (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164(4):805–817. doi:[10.1016/j.cell.2016.01.029](https://doi.org/10.1016/j.cell.2016.01.029)

iPTMnet: Integrative Bioinformatics for Studying PTM Networks

Karen E. Ross, Hongzhan Huang, Jia Ren, Cecilia N. Arighi, Gang Li, Catalina O. Tudor, Mengxi Lv, Jung-Youn Lee, Sheng-Chih Chen, K. Vijay-Shanker, and Cathy H. Wu

Abstract

Protein post-translational modification (PTM) is an essential cellular regulatory mechanism, and disruptions in PTM have been implicated in disease. PTMs are an active area of study in many fields, leading to a wealth of PTM information in the scientific literature. There is a need for user-friendly bioinformatics resources that capture PTM information from the literature and support analyses of PTMs and their functional consequences. This chapter describes the use of iPTMnet (<http://proteininformationresource.org/iPTMnet/>), a resource that integrates PTM information from text mining, curated databases, and ontologies and provides visualization tools for exploring PTM networks, PTM crosstalk, and PTM conservation across species. We present several PTM-related queries and demonstrate how they can be addressed using iPTMnet.

Key words Post-translational modification, Phosphorylation, Acetylation, Text mining, Protein-protein interaction, Protein ontology, Database, PTM crosstalk

1 Introduction

Post-translational modification (PTM) is a major mechanism by which the cell regulates the biological activity of proteins. PTMs, such as phosphorylation, acetylation, and ubiquitination, have a broad range of effects, altering protein stability, enzymatic activity, subcellular localization, and interactions. Coordination of multiple PTMs at the same site or at multiple sites on a protein affords another layer of complexity, giving the cell exquisite control over protein function [1]. Abnormalities in PTM have been implicated in many diseases, and modulation of PTM is being actively pursued as a therapeutic strategy. A growing number of kinase inhibitors are being used to treat cancer as well as inflammatory and autoimmune diseases [2]. Histone deacetylase inhibitors are also showing promise in cancer treatment [3].

Because of the extent and importance of PTM events in the cell, researchers from many fields are often confronted by PTM-related questions, from simple questions such as “What are the substrates of this kinase?” and “Which sites are acetylated in this protein?” to more complex queries such as “How does this PTM interact with other PTMs in the same protein?”, “What are the functional consequences of this PTM event?”, and “Which of these PTM events that have been observed in mouse are likely to also occur in humans?” The ultimate resource for answering these questions is the scientific literature; however, it can be overwhelming. A PubMed search for “phosphorylation” returns >250,000 articles; a search for “acetylation” returns nearly 30,000. Several efforts are underway to summarize this information in bioinformatics databases for easy consumption by biologists. Resources such as UniProtKB [4], PhosphoSitePlus [5], and Phospho.ELM [6] provide high-quality PTM information manually curated from the literature. However, manual curation is time- and labor-intensive, making it nearly impossible to keep up with the vast body of PTM literature. The use of automated text mining tools to capture PTM information from the literature is a promising approach to supplement the work of human curators.

We have developed iPTMnet (<http://proteininformationresource.org/iPTMnet/>), a user-friendly web resource that integrates text-mined information with information from curated databases to provide a detailed, current picture of PTM events. iPTMnet includes automated results from two text mining tools—RLIMS-P, which identifies mentions of kinases, substrates, and phosphorylation sites in text [7], and eFIP, which identifies mentions of phosphorylation-dependent protein-protein interactions (PPIs) [8]. The system incorporates PTM databases that specialize in different organisms, including mammals, plants, and yeast (*see Note 1*). With an emphasis on PTM relationships, including enzyme-substrate relationships and PTM-dependent PPIs, iPTMnet offers a user-customizable PTM network view. Furthermore, iPTMnet uses the Protein Ontology (PRO, *see Note 2*) [9] to represent combinatorial PTM forms (proteoforms, *see Note 3*) and orthologous relationships between PTM proteins in different organisms to support studies of PTM conservation across species and PTM crosstalk. Here we describe the use of iPTMnet to answer a variety of PTM-related biological questions (*see Note 4*).

2 Methods

2.1 Browsing in iPTMnet. Example: Overview of Plant Kinase Information

The iPTMnet Browse feature provides a general overview of PTMs and/or PTM enzymes in organisms of interest. To explore information about plant kinases:

1. Go to the iPTMnet home page at <http://proteininformationresource.org/iPTMnet/> and click “Browse” (Fig. 1).

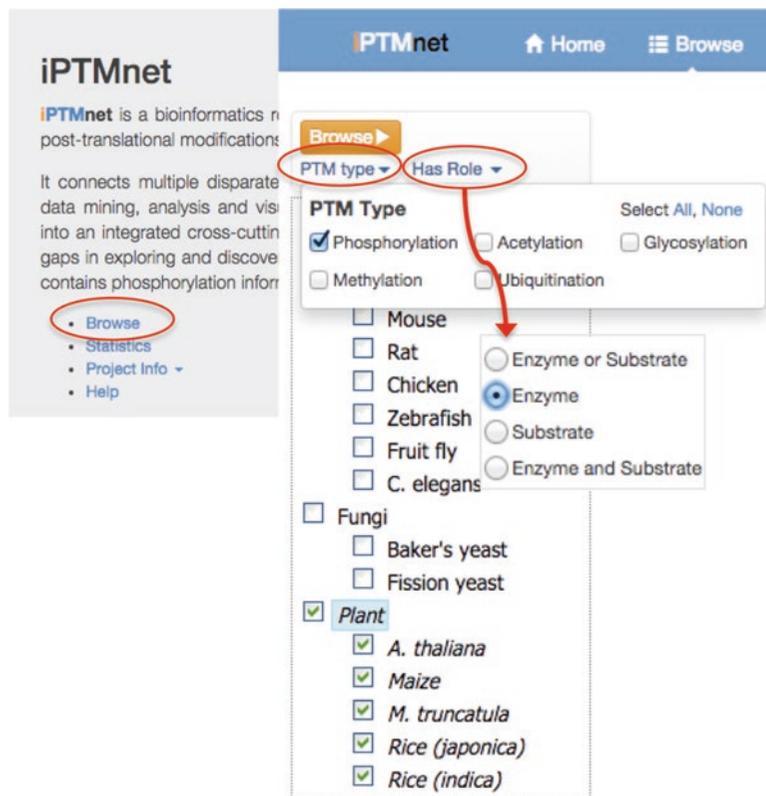


Fig. 1 Menu selections for browsing plant kinases in iPTMnet

- The panel on the left side of the page allows you to select individual organisms (e.g., Maize) or groups of organisms (e.g., Plant). For this example, select “Plant.” Output can be further filtered using the two menus—PTM type and Has Role—located directly below the orange “Browse” button. To display kinases, select “Phosphorylation” from the PTM type menu and “Enzyme” from the Has Role menu. Click “Browse” (Fig. 1).
- Results are displayed in a table with nine columns (Fig. 2). The first column shows the iPTMnet identifier. Clicking here will take you to the iPTMnet entry page for the protein (*see* Subheading 2.2). Clickable links to the iProClass, UniProt, and PRO pages for the protein are also provided. The next columns display the protein name, gene names, and organism. Reviewing the organism column reveals that most of the results (34/36) are proteins from *Arabidopsis*. The remaining two are from maize. The Substrate Role column displays a green check mark if the protein has known PTM sites and lists the number of known PTM enzymes, if any. Thirty-one out of the 36 plant kinases undergo PTM themselves (i.e., they have at least one PTM site); however, a PTM enzyme is known in only 17 of those cases. The Enzyme Role column displays a green check mark if the protein is a PTM enzyme, followed by the number

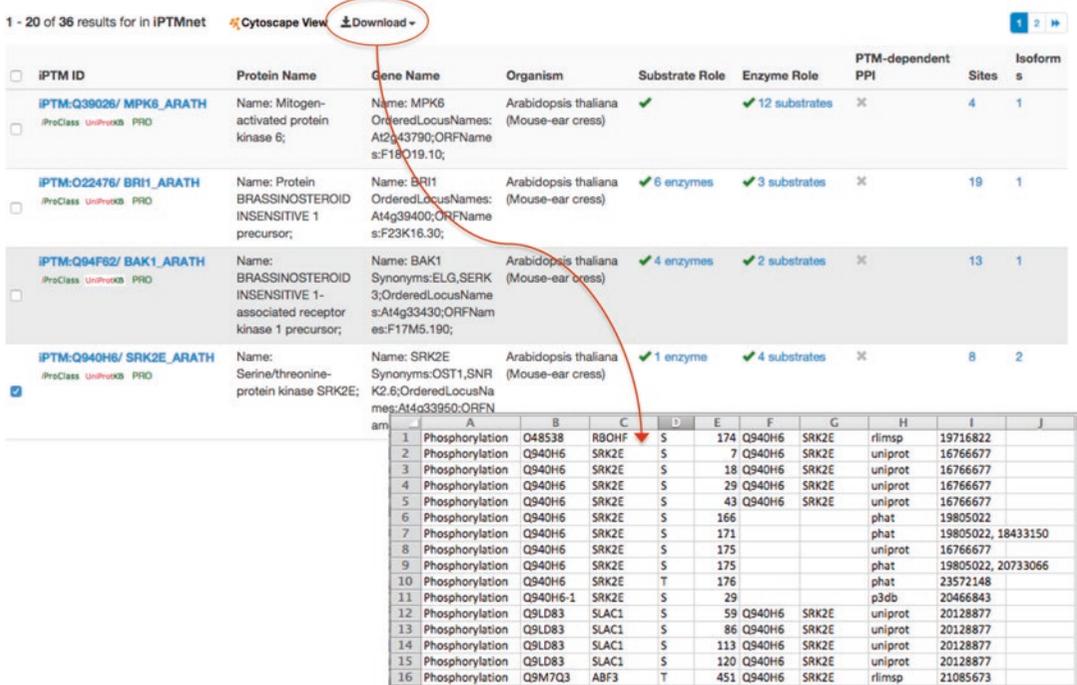


Fig. 2 Table of results of browsing plant kinases in iPTMnet. Users can select rows in the table and download the PTM information in tab-delimited format for viewing in a spreadsheet program such as Microsoft Excel

of known substrates. Because we filtered the list for proteins with kinase activity, every result has a green check mark in this column. The final three columns display the number of PTM-dependent PPIs automatically identified by eFIP, the number of PTM sites, and the number of isoforms for the protein. The numbers in the last five columns of the table are clickable links to the relevant sections of the protein entry page.

4. You can select results to download or display in a network view (*see* below) by checking the box to the left of the iPTMnet identifier (Fig. 2).

2.2 Basic Exploration of a PTM Network.
Example: Arabidopsis Mitogen-Activated Protein Kinase 6 (MPK6)

The iPTMnet protein entry pages are organized into several tables that provide information about PTM sites, PTM enzymes, proteoforms, and PTM-dependent PPIs for the selected protein. If the protein is itself a PTM enzyme, information about its substrates is also given. Users can access two visualization options from the entry pages: (1) a Cytoscape network view of the PTM relationships and (2) a multiple sequence alignment view that shows an alignment of proteoforms of the selected protein as well as orthologous proteins from other organisms with PTM sites highlighted (*see* Subheadings 2.5 and 2.6). We will illustrate how the protein entry

A Search for proteins in iPTMnet database

Search phosphon

Enter Keywords (accepts S

Input keyword

Exclude review papers

PTM type Has Role Restrict by Organism Sample Report

Select organisms

Human Mouse Rat

Cow Chicken Zebrafish

Fruit fly C. elegans Baker's yeast

Fission yeast A. thaliana Maize

M. truncatula Rice (aponica) Rice (indica)

Or input other organisms

Separate by comma

C iPTMnet Report for Q39026 (MPK6)

Protein Information

UniProt AC / UniProt ID Q39026 / MPK6_ARATH

Protein Name Name: Mitogen-activated protein kinase 6;

Gene Name Name: MPK6

OrderedLocusNames:At2g43790;ORFNames:F18O19.10;

Organization Name: Arabidopsis thaliana (Mouse-ear cress)

B 1 - 1 of 1 results for MPK6 in iPTMnet

iPTM ID	Protein Name	Gene Name	Organism	Substrate Role	Enzyme Role	PTM-dependent PPI	Sites	Isoforms
iPTM:Q39026/MPK6_ARATH	Name: Mitogen-activated protein kinase 6;	Name: MPK6 OrderedLocusNames:At2g43790;ORF Names:F18O19.10;	Arabidopsis thaliana (Mouse-ear cress)	✓	✓ 14 substrates	✗	3	1

Fig. 3 (a) The iPTMnet protein search interface. Settings for searching for MPK6 kinase in *Arabidopsis* are shown. (b) Results page for the search shown in (a). (c) Top portion of the iPTMnet entry page for *Arabidopsis* MPK6. The link to the Cytoscape network view for this protein is outlined in the green box

page can be used to answer PTM-related questions about a protein of interest using *Arabidopsis* MPK6 as an example. MPK6 is a MAPK, one of the terminal kinases in the MAPK signaling cascade, and plays a key role in the response to pathogens [10]. In general, MAPKs are activated via phosphorylation by upstream kinases. Therefore, we would expect MPK6 to be a phosphoprotein as well.

2.2.1 *Is MPK6 a Phosphoprotein? If So, What Sites Is It Phosphorylated on? What Phosphorylated Proteoforms of MPK6 Are Known?*

1. In the section labeled “Search for Protein in the iPTMnet Database” on the iPTMnet home page (<http://proteininformationresource.org/iPTMnet/>), enter MPK6 in the search box (Fig. 3a). Because we are interested in *Arabidopsis* MPK6, select *A. thaliana* from the “Restrict by Organism” menu, and click Submit.
2. Search results are displayed on a new page with the search term highlighted in yellow (Fig. 3b). The columns in the search results table are the same as in the browse results table described above. In this case, only one iPTMnet entry, our desired protein *Arabidopsis* MPK6, matches the search criteria. Click on the iPTMnet ID to go to the entry page (Fig. 3c).
3. To view information about PTM sites in MPK6, click on “Q39026 (MPK6) as Substrate” in the gray “Display” box in the upper left corner of the page (Fig. 4a). This will bring the substrate table to the top of the browser window (Fig. 4b). The table has five columns: Site, PTM Type, PTM Enzyme, Source, and PMID. We can see that MPK6 is indeed a phosphoprotein. It has four phosphorylation sites at S215, T221,

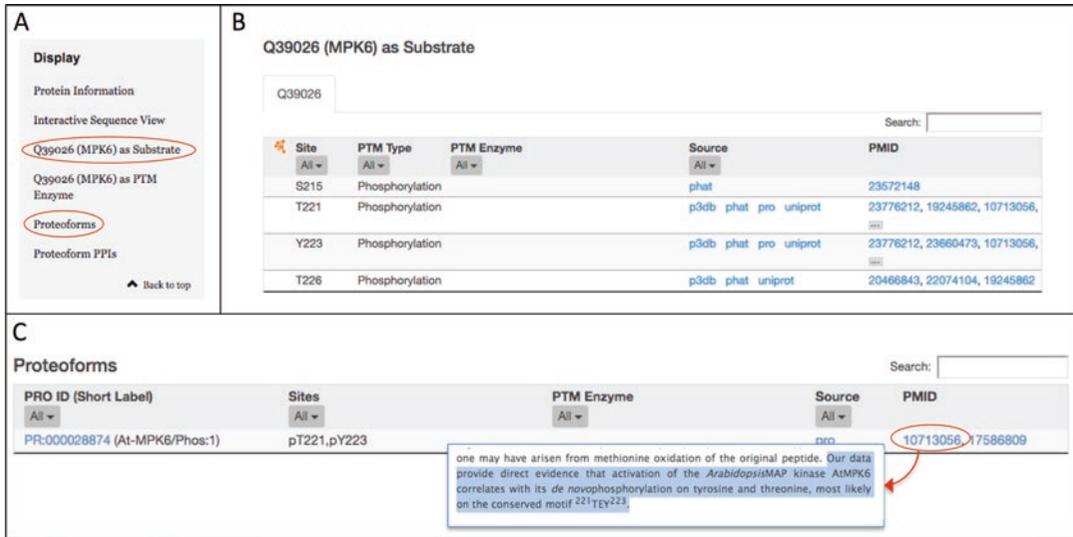


Fig. 4 Parts of the iPTMnet entry page for *Arabidopsis* MPK6. (a) The Display panel with links to the various tables on the entry page. Links to the MPK6 as Substrate table shown in (b) and the Proteoforms table shown in (c) are indicated by the red circles. (b) The MPK6 as Substrate table, which shows information about PTM sites on MPK6. (c) Proteoforms table, which shows PTM proteoforms of MPK6 that have been curated in PRO. Clicking on links in the PMID column (red circle) will display the PubMed abstracts of articles used as evidence for the PRO terms. From there, users can navigate to the full-text articles, if available, using the PubMed interface. An evidence sentence taken from the Results section of PMID:10713056 is shown in the blue box

Y223, and T226. The PTM enzyme is not known for any of these sites. We can view the information in its source database or view the associated literature citations, by clicking on the links in the Source and PMID columns, respectively.

- To view proteoforms of MPK6, click on “Proteoforms” in the “Display” box (Fig. 4c). The Proteoform table lists the proteoform ID, the site and type of modification, the modifying enzymes if known, source, and literature reference. One MPK6 proteoform has been described: MPK6/Phos:1 (PR:000028874), which is doubly phosphorylated on T221 and Y223, two of the sites listed in the Substrate table. Inspection of the Results section of one of the literature references by clicking on its PMID, 10713056, and then navigating to the full-text article via PubMed indicates that this proteoform is the active form of MPK6 [11].

2.2.2 What Are the Substrates of MPK6?

To view substrates of MPK6, click on “Q39026 (MPK6) as PTM Enzyme” in the “Display” box. The PTM enzyme table has two tabs. The first tab “Protein as Phosphorylation Enzyme” (Fig. 5a) shows all substrates of MPK6 in cases where the relevant MPK6 proteoform is not reported; the second tab “Proteoform as Phosphorylation Enzyme” (Fig. 5b) shows the substrates of specific proteoforms of MPK6. The “Protein as Phosphorylation

Q39026 (MPK6) as PTM Enzyme A

Protein as Phosphorylation Enzyme **Proteome as Phosphorylation Enzyme**

Search:

Substrate	Site	Source	PMID
All ▾	All ▾	All ▾	
<input type="checkbox"/> O22533 (ZAT6)	S8	rimsp	23257164
<input type="checkbox"/> O22533 (ZAT6)	S223	rimsp	23257164
<input type="checkbox"/> O80982 (HSFA2)	T249	rimsp	23638397
<input type="checkbox"/> P11035 (NIA2)	S627	rimsp	21593598
<input type="checkbox"/> Q8VZ91 (ERF6)	S266	pro	23300166
<input type="checkbox"/> Q8VZ91 (ERF6)	S269	pro	23300166
<input type="checkbox"/> Q9FDW1 (MYB44)	S53	rimsp	22704933
<input type="checkbox"/> Q9FDW1 (MYB44)	S145	rimsp	22704933
<input type="checkbox"/> Q9M0J5 (MYB41)	S251	pro rimsp	22575450
<input type="checkbox"/> Q9SJF3	S237	rimsp	22407295

Q39026 (MPK6) as PTM Enzyme B

Protein as Phosphorylation Enzyme **Proteome as Phosphorylation Enzyme**

Search:

PRO ID (Short Label)	Sites	PTM Enzyme	Source	PMID
All ▾	All ▾	All ▾	All ▾	
PR:000037425 (At-EIN3/Phos:2)	pT174	PR:000028874 (At-MPK6/Phos:1)	pro	18273012
PR:000028859 (At-ACS2/Phos:1)	pS483,pS488,pS491	PR:000028874 (At-MPK6/Phos:1)	pro	15539472, 18182027, 20659280
PR:000028864 (At-ERF104/MAPK-PhosRes-)		PR:000028874 (At-MPK6/Phos:1)	pro	19416906
PR:000028854 (At-ACS6/Phos:1)	pS480,pS483,pS488	PR:000028874 (At-MPK6/Phos:1)	pro	15539472
PR:000037415 (At-DCP1A/Phos:1)	pS237	PR:000028874 (At-MPK6/Phos:1)	pro	22407295

Fig. 5 The MPK6 as PTM enzyme table for *Arabidopsis* MPK6. The table has two tabs. The first tab (a) shows substrate and site information for MPK6 in general. The second tab (b) shows substrate and site information for a phosphorylated proteoform of MPK6 (At-MPK6/Phos:1, PR:000028874)

Enzyme” table lists the ID of the substrate, the site, the source database, and literature references (*see Note 5*). Each substrate-site pair is listed in its own row. Thus, the first and second rows of the table indicate that MPK6 phosphorylates the substrate ZAP6 on S8 and S223, respectively. Overall, there are seven substrates, and ten substrate-site pairs listed. The “Proteome as Phosphorylation Enzyme” table lists the ID of the modified proteoform, the modification sites, and the ID of the MPK6 proteoform that is acting as the PTM enzyme as well as source information and references. In this table we see that the doubly phosphorylated activated form of MPK6 (MPK6/Phos:1, PR:000028874) phosphorylates one or more sites on five different proteins (*see Note 6*).

2.2.3 Visualization of the MPK6 Phosphorylation Network

For a network view of MPK6 relationships, click on the Cytoscape View icon at the top right of the page in the Protein Information section (green box in Fig. 3c). The network will be displayed in a new tab/window (Fig. 6). Click on “LEGEND” in the upper left corner of the window to display an explanation of the node and

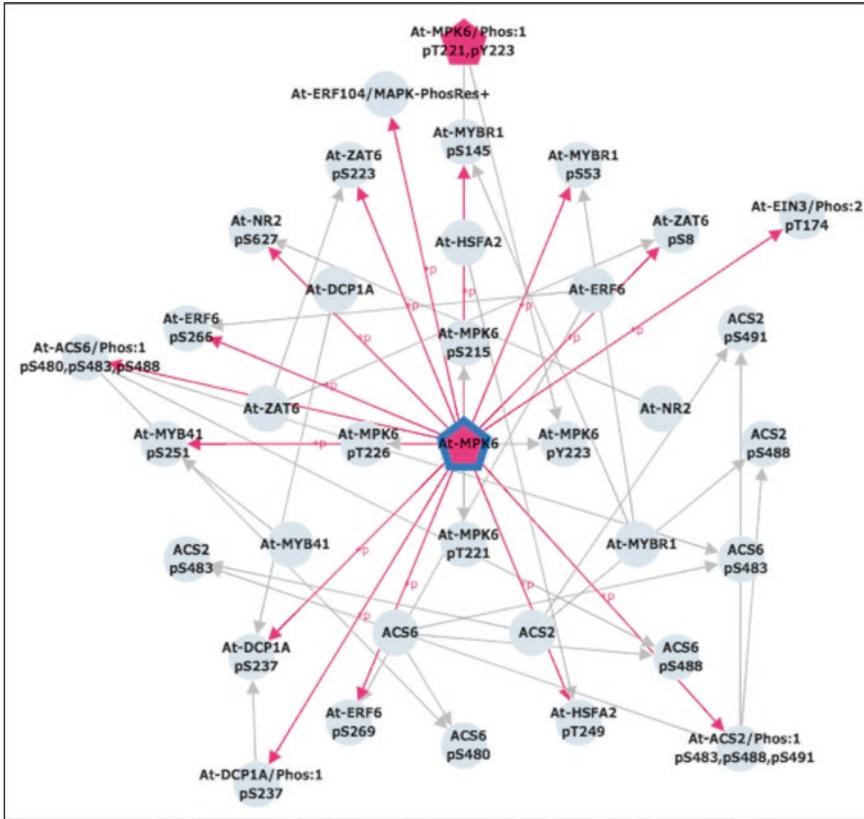


Fig. 6 Cytoscape network view for *Arabidopsis* MPK6. Kinases are pink, pentagonal nodes; proteins, sites, and proteoforms are gray circles. Gray edges connect sites to their corresponding proteins and proteoforms. Pink edges connect kinases to the sites they phosphorylate. MPK6 itself is outlined in blue

edge style. PTM enzyme nodes are pentagons and substrates and sites are circles. PTM-enzyme → site, PTM-enzyme → proteoform and site → substrate edges are shown. The color of the PTM-enzyme nodes and PTM-enzyme → site/proteoform edges indicates the PTM type (e.g., pink represents phosphorylation).

2.3 PMID-Centric Searching in iPTMnet

iPTMnet also allows users to search for a PMID of interest and view a summary of all PTM information in iPTMnet associated with that PMID.

1. From the iPTMnet home page, select “PMID” from the pull-down menu to the left of the “Search for Proteins in iPTMnet Database” panel. Enter a PMID of your choice, for example, 15696159, and click Submit (Fig. 7a).
2. Inspect the results page. The page shows the PMID, title, and abstract of the article (Fig. 7b), followed by several tables of PTM information (Fig. 7c–f). The selection of tables that is

Search for proteins in iPTMnet database

PMID: 15696159 **Submit** **Reset**

PTM type: Has Role Restrict by Organism [Sample Report](#)

PMID: 15696159

JNK phosphorylation of 14-3-3 proteins regulates nuclear targeting of c-Abl in the apoptotic response to DNA damage.

Abstract

The ubiquitously expressed c-Abl tyrosine kinase localizes to the cytoplasm and nucleus. Nuclear c-Abl is activated by diverse genotoxic agents and induces apoptosis; however, the mechanisms that are responsible for nuclear targeting of c-Abl remain unclear. Here, we show that cytoplasmic c-Abl is targeted to the nucleus in the DNA damage response. The results show that c-Abl is sequestered into the cytoplasm by binding to 14-3-3 proteins. Phosphorylation of c-Abl on Thr 735 functions as a site for direct binding to 14-3-3 proteins. We also show that, in response to DNA damage, activation of the c-Jun N-terminal kinase (Jnk) induces phosphorylation of 14-3-3 proteins and their release from c-Abl. Together with these results, expression of an unphosphorylated 14-3-3 mutant attenuates DNA-damage-induced nuclear import of c-Abl and apoptosis. These findings indicate that 14-3-3 proteins are pivotal regulators of intracellular c-Abl localization and of the apoptotic response to genotoxic stress.

PTM Type	Substrate	Site	PTM Enzyme	Source
Phosphorylation	iPTM:P00519-1 (ABL1) :ProClass UniProtKB PRD	T735		hypr pro rimsp uniprot
Phosphorylation	iPTM:P31946 (YWHAB) :ProClass UniProtKB PRD	S186	iPTM:P45983 (MAPK8) :ProClass UniProtKB PRD	peim psp

Proteoforms

PRO ID (Short Label)	Sites	PTM Enzyme	Source	PMID
PR:000044506 (tyrosine-protein kinase ABL1 phosphorylated 1 (human))	pT735		pro	15696159, 16888623
PR:000044508 (14-3-3 protein zeta/delta phosphorylated 1 (human))	pS184		pro	15696159

Proteoform PPIs

Protein 1	Relation	Protein 2	Source
PR:000044506 (tyrosine-protein kinase ABL1 phosphorylated 1 (human))	Interaction	PR:000044507 (14-3-3 protein (human))	pro
PR:000044506 (tyrosine-protein kinase ABL1 phosphorylated 1 (human))	Interaction	PR:P63104 (hYWHAZ)	pro

PTM-dependent PPI

PTM type	Substrate	Site	Interactant	Association type	Source
Phosphorylation	iPTM:P00519 (ABL1) :ProClass UniProtKB PRD	T735	iPTM:P27348 (YWHAQ) :ProClass UniProtKB PRD	unknown association	efip

Fig. 7 PMID-centric view in iPTMnet. Searching for PMID:15696159 (a) opens the PMID:15696159 entry page, which displays the title and abstract of the article (b) as well as several tables of iPTMnet information associated with the article: a PTM enzyme-substrate table (c), a proteoform table (d), a proteoform-PPI table showing curated phosphorylation-dependent PPIs from PRO (e), and a PTM-dependent PPI table showing phosphorylation-dependent PPIs automatically extracted by eFIP (f)

displayed will depend on the PTM information in the article. In this case, there are four tables:

Enzyme-substrate table (Fig. 7c): This table summarizes the PTM enzyme, substrate, and site information in iPTMnet for which the article is cited as evidence. We can see that this article described phosphorylation of ABL1 and 14-3-3 proteins (YWHAB and YWHAZ). In two cases (YWHAB pS186 and YWHAZ pS184), the kinase, MAPK8, is also provided.

Proteoform table (Fig. 7d): This table displays the proteoforms curated by PRO based on the article. Two of the substrate-site pairs shown in the first table have been curated by PRO: ABL1 pT735 (PR:000044506) and 14-3-3 protein zeta/delta (YWHAZ) pS184 (PR:000044508).

Proteoform PPIs (Fig. 7e): This table shows the phosphorylation-dependent PPIs that have been curated by PRO based on the article. For example, the tyrosine-phosphorylated form of ABL1 (PR:0000544506) interacts with YWHAZ.

PTM-dependent PPI (Fig. 7f): This table displays phosphorylation-dependent PPIs automatically extracted from the article by the text mining tool, eFIP. In this case, eFIP detected an interaction between phosphorylated ABL1 and the 14-3-3 family member, YWHAQ.

2.4 Construction and Analysis of More Complex PTM Networks. Example: Tyrosine Phosphorylation of Beta-Catenin (CTNNB1) During Mitosis

Users can construct networks based on multiple iPTMnet entry pages in order to address more complex scientific questions as shown in the following example involving beta-catenin (CTNNB1). CTNNB1 functions as an adhesion molecule as part of the adherens junction at the cell membrane and as a transcriptional co-regulator in the nucleus. The distribution of CTNNB1 between the membrane-associated and nuclear pools and CTNNB1 stability are regulated by a complex interplay of multiple PTMs. Tyrosine phosphorylation of CTNNB1, which occurs on several residues, is generally associated with CTNNB1 dissociation from the membrane, increased stability, and increased transcriptional activity [12]. It has been reported that CTNNB1 tyrosine phosphorylation decreases during mitosis [13]. It is plausible that this decrease is due to regulation of one or more of the CTNNB1 tyrosine kinases by the mitotic kinase CDK1. We can use iPTMnet to identify tyrosine kinases that phosphorylate CTNNB1 that are in turn phosphorylated by CDK1:

1. Search for the CTNNB1 entry page in iPTMnet as in **step 1** of Subheading 2.2.1. Enter “CTNNB1” in the search box and select human as the organism. You will see one search result; click on its iPTMnet ID (iPTM:P35222/CTNNB1_HUMAN) to go to the entry page.
2. Go to the “P35222 (CTNNB1) as Substrate” table. Because we are interested in CTNNB1 tyrosine phosphorylation, use the pull-down menu in the Site column to select “All Tyrosine” and the pull-down menu in the PTM type column to select “Phosphorylation.” This will filter the table to show only tyrosine phosphorylation events (Fig. 8).
3. Next, we will build a custom network view that only displays the kinase-site relations for the tyrosine phosphorylation sites. For each site that has a known kinase, there is a check box to the left of the site column (Fig. 8, red arrow). Click on all of the check boxes in the filtered Substrate table. As you click on each one, the substrate ID, site, and kinase ID will appear in the gray Cytoscape View panel on the left. Click on the submit button at the bottom of the Cytoscape View panel (Fig. 8, green rectangle). The network will be displayed in a new tab/window. You can manually position the nodes by clicking and dragging to make a concentric layout with CTNNB1 in the center, its tyrosine phosphorylation sites in the inner ring, and

Cytoscape View Clear

P35222/pY654-P00519 ✕

P35222/pY654-P12931 ✕

P35222/pY654-P41240 ✕

P35222/pY654-P00533 ✕

P35222/pY654-P36888 ✕

P35222/pY489-P00519 ✕

P35222/pY333-P41240 ✕

P35222/pY333-Q13882 ✕

P35222/pY333-P12931 ✕

P35222/pY331-P41240 ✕

P35222/pY331-Q13882 ✕

P35222/pY142-P06241 ✕

P35222/pY142-P41240 ✕

P35222/pY142-P22607 ✕

P35222/pY142-Q13882 ✕

P35222/pY86-P00519 ✕

P35222/pY86-P41240 ✕

P35222/pY64-Q13882 ✕

P35222/pY64-P41240 ✕

Site	PTM Type	PTM Enzyme	Source
All Tyrosine	Phosphorylation	All	All
<input type="checkbox"/> Select all	<input type="checkbox"/> Select all		psp
<input type="checkbox"/> By amino acid	<input type="checkbox"/> Acetylation	2 (PTK6)	psp rimsp uniprot
<input type="checkbox"/> All Serine	<input checked="" type="checkbox"/> Phosphorylation	0 (CSK)	uniprot
<input type="checkbox"/> All Threonine	<input type="checkbox"/> Ubiquitination	9 (ABL1)	psp
<input checked="" type="checkbox"/> All Tyrosine	Phosphorylation	Q13882 (PTK6)	psp rimsp uniprot
<input type="checkbox"/> By position	Phosphorylation	P04629 (NTRK1)	psp
<input type="checkbox"/> K19	Phosphorylation	P21802 (FGFR2)	psp
<input type="checkbox"/> Y142	Phosphorylation		rimsp
<input checked="" type="checkbox"/> Y142	Phosphorylation	P22607 (FGFR3)	psp rimsp
<input checked="" type="checkbox"/> Y142	Phosphorylation	P00533 (EGFR)	psp
<input checked="" type="checkbox"/> Y142	Phosphorylation	P06241 (FYN)	hprd palm
<input type="checkbox"/> Y331	Phosphorylation		rimsp
<input checked="" type="checkbox"/> Y331	Phosphorylation	Q13882 (PTK6)	psp uniprot
<input checked="" type="checkbox"/> Y333	Phosphorylation	P12931 (SRC)	psp
<input type="checkbox"/> Y333	Phosphorylation		pro rimsp
<input checked="" type="checkbox"/> Y333	Phosphorylation	Q13882 (PTK6)	psp uniprot
<input checked="" type="checkbox"/> Y489	Phosphorylation	P00519 (ABL1)	psp rimsp
<input checked="" type="checkbox"/> Y654	Phosphorylation	P36888 (FLT3)	rimsp
<input checked="" type="checkbox"/> Y654	Phosphorylation	P00533 (EGFR)	pro psp
<input checked="" type="checkbox"/> Y654	Phosphorylation	P41240 (CSK)	uniprot
<input checked="" type="checkbox"/> Y654	Phosphorylation	P00519 (ABL1)	psp

Fig. 8 Portion of the human CTNNB1 entry page illustrating how to construct a custom Cytoscape network view of tyrosine phosphorylation of CTNNB1. The *right side* of the figure shows the settings for filtering the CTNNB1 as substrate table to show only tyrosine phosphorylation sites. Clicking the *checkboxes* next to the tyrosine sites (*red arrow*) adds these relations to the custom Cytoscape View panel (*left*). Clicking on the Submit button (outlined in *green*) will display the network

the kinases in the outer ring (Fig. 9). The network shows seven CTNNB1 tyrosine phosphorylation sites (Y64, Y86, Y142, Y331, Y333, Y489, Y654) that are phosphorylated by ten kinases (EGFR, FLT3, FYN, FGFR2, FGFR3, PTK6, CSK, SRC, NTRK1, and ABL1; the “h” preceding each kinase name in the Cytoscape network indicates that they are from human).

- The next step is to determine which of these tyrosine kinases is a substrate of the mitotic kinase CDK1. Return to the tab/window displaying the CTNNB1 entry page. Do not clear the Cytoscape View panel. Click on the iPTMnet logo in the upper left corner of the page to go to the home page. Search for CDK1 in human as you did for CTNNB1 in **step 1** of this Subheading. Click on the search result (iPTM:P06493/CDK1_HUMAN) to go to the human CDK1 entry page.
- Go to the “P06493 (CDK1) as PTM Enzyme” table. Enter the first CTNNB1 tyrosine kinase from **step 3** (EGFR) into the search box in the upper right corner of the table (Fig. 10a). The table will display only one row, which indicates that EGFR S1026 is phosphorylated by CDK1. Check the box to the left of the site column to add this relation to the custom Cytoscape view.

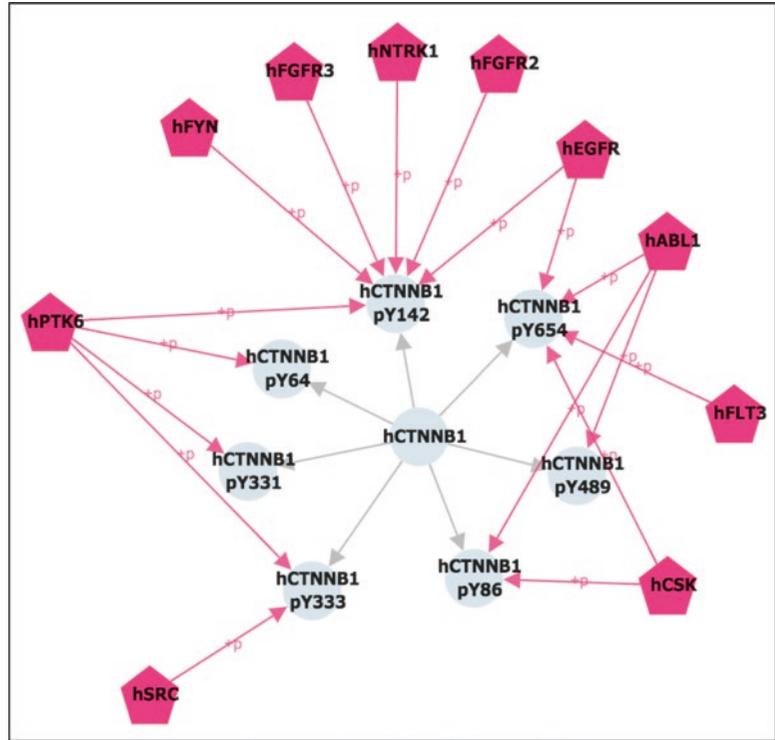


Fig. 9 Cytoscape view of tyrosine phosphorylation of CTNNB1. Kinases are pink, pentagonal nodes; CTNNB1 and its phosphorylation sites are *gray circles*. *Gray edges* connect sites to CTNNB1. *Pink edges* connect kinases to the sites they phosphorylate

The Cytoscape panel should still be displaying the CTNNB1 kinase-substrate relations added in **step 3**. Next, search for the second kinase, FLT3, in the table. iPTMnet does not have any information about CDK1 phosphorylation of FLT3, so the search returns, “No matching records found” (Fig. 10b). Continue to search for the remaining nine kinases. Of these kinases, only one (ABL1) is a CDK1 substrate; add this relation to the Cytoscape view. Finally, click on the submit button in the Cytoscape View panel to see the network.

6. You can manually position the nodes to create the arrangement shown in Fig. 11. The gray outline shows the CDK1 relations that were added to the network. From this analysis, we can conclude that CDK1 may regulate EGFR through phosphorylation on S1026 and ABL through phosphorylation on S569, which, in turn, could affect CTNNB1 phosphorylation on Y86, Y142, Y489, and Y654. To test the hypothesis that CDK1 is negatively regulating CTNNB1 phosphorylation via EGFR and/or ABL, the next step would be to check the literature references for the CDK1-EGFR and CDK1-ABL kinase-substrate relations to see what effect CDK1 phosphorylation has on EGFR and ABL activity.

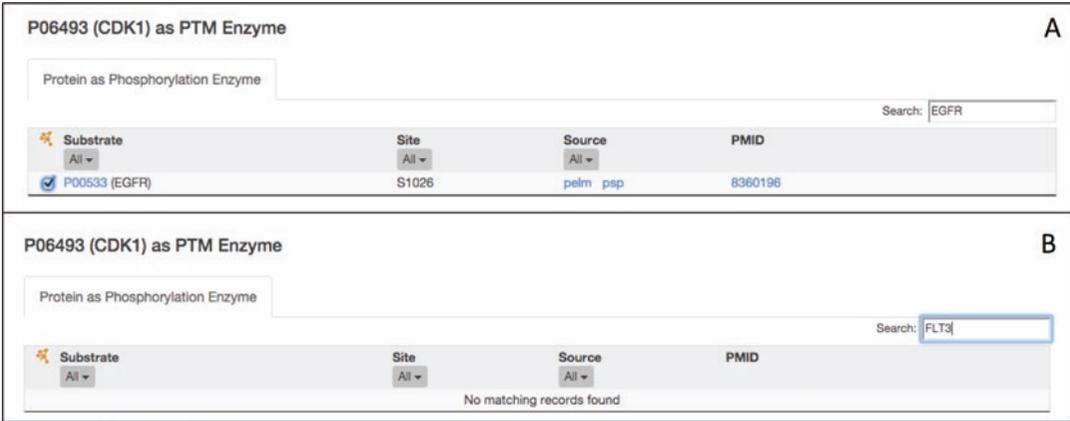


Fig. 10 Views of the human CDK1 as PTM Enzyme table, illustrating how to search for particular substrates of interest. Searching the table for EGFR (a) returns one result, indicating that CDK1 phosphorylates EGFR on one site (S1026). Searching for FLT3 (b) returns no results, indicating that iPTMnet does not have any information about FLT3 phosphorylation by CDK1

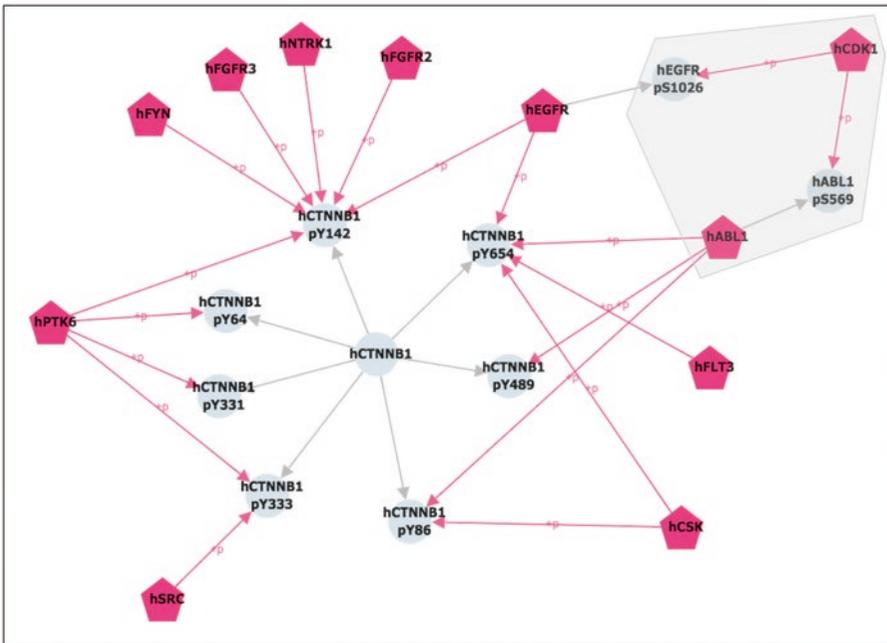


Fig. 11 Cytoscape network view showing potential regulation of CTNNB1 tyrosine phosphorylation by CDK1. The CTNNB1 tyrosine phosphorylation network is shown as in Fig. 9. In addition, phosphorylation of CTNNB1 tyrosine kinases by CDK1 is shown (gray outline). Kinases are pink, pentagonal nodes; other proteins and their phosphorylation sites are gray circles. Gray edges connect sites to their corresponding proteins. Pink edges connect kinases to the sites they phosphorylate

2.5 Analysis of PTM Crosstalk in iPTMnet.

Example: TP53 and EP300

PTM crosstalk is the influence of one PTM on other PTMs of the same substrate protein. Crosstalk can involve direct competition between two PTMs for a single site, or it can occur when PTM at one site enhances or inhibits binding of PTM enzymes, thereby influencing PTM at a second site. iPTMnet can be used to explore both of these types of crosstalk. To facilitate identification of cases where two PTMs compete for the same site, sites that undergo multiple types of PTM are highlighted in a special dark yellow color in the iPTMnet sequence view. To identify potential crosstalk among multiple sites in a protein, we can take advantage of the integration of PTM-dependent PPI and PTM-enzyme-substrate information in iPTMnet. Specifically, we can find cases where a PTM enzyme is involved in a PTM-dependent PPI with a protein via one modification site and also modifies a second site on the same protein. Although the existence of these multiple relations does not definitively show that PTM crosstalk is occurring, it pinpoints interesting candidates for further study. In this Subheading, we will illustrate how to use iPTMnet to explore PTM crosstalk using the tumor suppressor protein TP53 and the acetyltransferase EP300 as examples.

2.5.1 PTM Crosstalk I: Exploring Sites with Multiple Possible Modifications

1. Using a similar procedure to **steps 1** and **2** of Subheading [2.2.1](#), go to the iPTMnet entry page for human TP53. Enter “TP53” in the search box and select human from the Restrict by Organism menu.
2. Look at the Interactive Sequence View Panel. It may take a few extra seconds to load. You will see a “zoomed-out” view of the human TP53 protein sequence. Only the modified residues are shown, highlighted in different colors. The color of the highlighting indicates the type of modification: pink is phosphorylation, light blue is methylation, dark blue is acetylation, and dark yellow is multiple modifications. You can scroll through the sequence using the horizontal scroll bar.
3. Click on the magnifying glass in the upper right corner of the panel. Now you will see a “zoomed-in” view of the sequence that displays all residues. The modified residues are still highlighted as before and you can scroll through the sequence using the horizontal scroll bar (Fig. [12a](#)). Clicking on the magnifying glass again will return you to the “zoomed-out” view.
4. Moving the mouse over a highlighted residue in either the “zoomed-out” or “zoomed-in” view will open a box with information about the modification, including modification type, PTM enzyme (if known), evidence source, and literature references. To examine this information for a multiply modified site that may serve a point of PTM crosstalk, mouse over the dark yellow highlighted serine residue at position 149 (Fig. [12a](#)). The box reveals that this site is both phosphorylated by casein

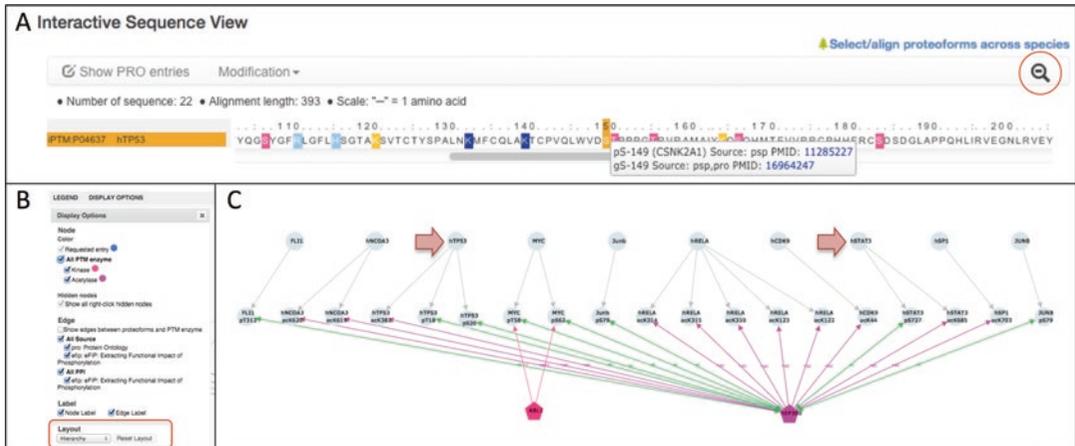


Fig. 12 PTM crosstalk involving human TP53 and EP300. **(a)** Interactive sequence view panel on the human TP53 entry page. Modified sites are highlighted (*pink*, phosphorylation; *light blue*, methylation; *dark blue*, acetylation; *dark yellow*, multiple modifications). Mousing-over a modified site (e.g., S149) pops up a window with more information about the modification. Clicking on the magnifying glass (*red circle*) toggles between a zoomed-out and zoomed-in view of the sequence. The zoomed-in view is shown. **(b)** Display Options panel in the Cytoscape network view window showing network style parameters that can be set by the user. The menu for changing the network layout is indicated in the *red rectangle*. **(c)** Cytoscape network view showing PTM-dependent PPIs (*green edges*) and PTM enzyme-site (*purple edges*) relations for human EP300. The *arrows* indicate proteins that are involved in phosphorylation-dependent PPIs with EP300 and are also acetylated by EP300. PTM enzymes are pentagonal nodes (*pink*, kinases; *purple*, acetylases). Other proteins and sites are *gray circles*. *Gray edges* connect sites to their corresponding proteins; *pink edges* indicate phosphorylation; *purple edges* indicate acetylation; and *green edges* indicate phosphorylation-dependent PPIs

kinase II (CSNK2A1) and glycosylated. Click on the PMIDs to view the evidence for these modifications and learn more about how they may interact.

2.5.2 PTM Crosstalk II: Exploring Crosstalk Among Multiple Residues

1. Using a similar procedure to **steps 1** and **2** of Subheading **2.2.1**, go to the iPTMnet entry page for the human acetyltransferase EP300. Enter “EP300” in the search box and select human from the Restrict by Organism menu. Click on the search result (iPTM:Q09472/ EP300_HUMAN) to go to the entry page.
2. Inspect the “Q09472 (EP300) as PTM Enzyme” table to view substrates/sites that are acetylated by EP300. One of the substrates listed is TP53, which is acetylated by EP300 on K382. Another, STAT3, is acetylated on K685.
3. Next, inspect the PTM-dependent PPI table (the last table on the entry page). You will see that EP300 (the interactant) participates in several PTM-dependent PPIs, including one with TP53 phosphorylated on T18 or S20. Thus, EP300 participates in a PTM-dependent PPI with TP53 and also acetylates TP53, raising the possibility that there is crosstalk between

these two modifications. To learn more about the nature of this potential crosstalk, click on the PMID for the TP53-EP300 interaction (PMID:11258706) and review the abstract. The abstract states that phosphorylation of EP300 on T18 and S20 enhances EP300 binding, suggesting that phosphorylation of TP53 on T18 and S20 may lead to increased acetylation at K382 by EP300. Similarly, EP300 participates in PTM-dependent PPI with STAT3 phosphorylated on S727. Since STAT3 is also acetylated by EP300, there is also possible crosstalk between these modifications.

4. This analysis can also be performed using the Cytoscape network view. To create the view, first click on “Clear” in the Cytoscape View panel to remove any previous selections. Next, go to the “Q09472 (EP300) as PTM Enzyme” table and click the checkboxes next to each EP300 substrate/site pair. These relations should appear in the Cytoscape View panel. Then, go to the PTM-dependent PPI table and click on all of the checkboxes to add these relations to the network view. Click on “Submit” in the Cytoscape View panel, which will display the network in a new tab. To more easily identify cases where proteins participate in multiple types of relations with EP300 (i.e., PTM-enzyme-site and PTM-dependent PPI), open the Display Options menu and select “Hierarchical” from the Layout menu (Fig. 12b). Inspect the network (Fig. 12c). You should see that TP53 has three modification sites that are involved in relations with EP300: two phosphorylation sites (T18 and S20) that participate PTM-dependent PPIs (green edges) and one site (K382) that is acetylated. STAT3 has two sites involved in relations with EP300: one phosphorylation site (S727) that participates in a PTM-dependent PPI and one acetylation site (K685). The other proteins in the network are involved in either PTM-dependent PPI or acetylation relations with EP300, but not both.

2.6 Exploration of Cross-Species Conservation of PTMs in iPTMnet. Example: BCL2 in Human, Mouse, and Rat

The iPTMnet Interactive Sequence Alignment view provides a convenient interface for comparing PTM proteoforms within and across species. The view uses the PRO hierarchy to identify related sequences to align. Experimentally observed PTM sites in each sequence are highlighted in color. If the modifiable residue is conserved in other sequences in the alignment, it is highlighted in gray. Users can select which sequences and/or PTMs to show in the alignment. We will use the iPTMnet sequence alignment view to compare phosphorylation of the anti-apoptotic and cell cycle inhibitory protein BCL2 in human, mouse, and rat:

1. Go to the iPTMnet entry page for human BCL2 (P10415).
2. In the upper right corner of the Interactive Sequence View panel, click on “Select/align proteoforms across species”.

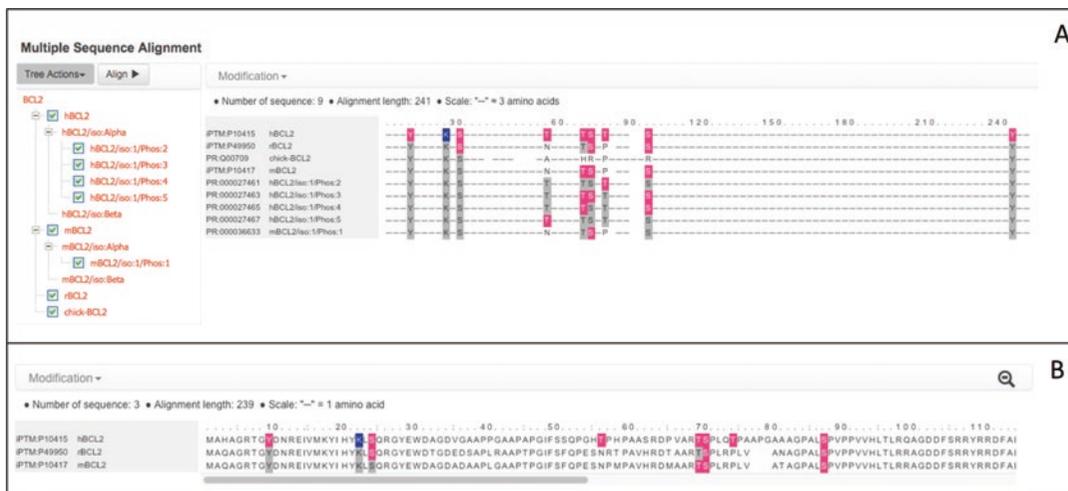


Fig. 13 (a) Multiple sequence alignment view for BCL2. Phosphorylated sites are indicated in *pink*; conserved phosphorylation sites are indicated in *gray*. The zoomed-out sequence is shown. On the *left*, the tree view for selecting sequences to include in the alignment is shown. **(B)** Zoomed-in multiple sequence alignment view for three selected BCL2 sequences: human (hBCL2), rat (rBCL2), and mouse (mBCL2)

The alignment will open in a new tab (Fig. 13a). Nine sequences are shown by default. The first four sequences are BCL2 orthologs from different species (human/hBCL2, rat/rBCL2, chicken/chick-BCL2, and mouse/mBCL2). Human BCL2 has eight phosphorylation sites (highlighted in pink) and one ubiquitination site (highlighted in dark blue). One human phosphorylation site, Y9, is conserved in mouse, rat, and chicken (as indicated by the gray highlighting) but has not been experimentally shown to be phosphorylated in these organisms. Another site, S70, has been shown to be phosphorylated in human, mouse, and rat, but is not conserved in chicken. The remaining sequences are proteoforms of BCL2 (four human and one mouse) with experimentally observed combinations of phosphorylation sites. For example, hBCL2/iso:1/Phos:3 (PR:000027463) is phosphorylated on three of the six possible sites (T69, S70, and S87). hBCL2/iso:1/Phos:4 (PR:000027465) is phosphorylated on T69 and S87, but not S70.

- At the left of the alignment, there is a hierarchical tree view that lists the sequences shown in the alignment (Fig. 13a). To more easily compare human, mouse, and rat BCL2, uncheck all of the sequences in the tree except hBCL2, mBCL2, and rBCL2 and click on “Align.” Click on the magnifying glass in the upper right corner of the alignment to see the “zoomed-in” view (Fig. 13b). Six of the human phosphorylation sites (Y9, S24, S69, S70, S87, and Y235) and the ubiquitination site (K22) are conserved in all three organisms. In addition, S70

and S87 have been shown to be phosphorylated in all three organisms. However, the remaining two sites, T56 and T74, are not conserved in rat or mouse.

4. Mouse-over human T56 and T74 to see the kinase information and references for these sites. T56 is phosphorylated by MAPK1, MAPK3, MAPK14, and CDK1; T74 is phosphorylated by MAPK1 and MAPK3. Click on the PMIDs to learn more about the regulation and functional consequences of these modifications. According to PMID:10669763, phosphorylation of T56, T74, and S87 by MAPK family members protects BCL2 from degradation, thereby preserving its anti-apoptotic activity. S87, which is phosphorylated in human, mouse, and rat, appears to be the most important site for preventing degradation. Thus, it is possible that this regulatory mechanism is conserved in mouse and rat and mediated by phosphorylation S87 only [14]. However, according to PMID:10766756, phosphorylation of T56 by CDK1 is required for the cell cycle inhibitory function of BCL2 [15]. Lack of conservation of this site in mouse and rat raises interesting questions about the cell cycle role of BCL2 in these organisms.

3 Notes

1. The following PTM resources are included in iPTMnet:
 - (a) HPRD: post-translational modifications and enzyme-substrate relationships for human proteins [16]
 - (b) PhosphoSitePlus: expert-curated PTM information including phosphorylation, ubiquitination, acetylation, and methylation mainly for human, rat, and mouse proteins [5]
 - (c) Phospho.ELM: expert-curated database for phosphorylation sites in animal proteins [6]
 - (d) PhosPhAt: protein phosphorylation sites identified by mass spectrometry in *Arabidopsis thaliana* [17]
 - (e) P3DB: protein phosphorylation data from multiple plants derived from large-scale experiments and the literature [18]
 - (f) PhosphoGrid: experimentally verified in vivo protein phosphorylation sites in *Saccharomyces cerevisiae* [19]
 - (g) UniProtKB: comprehensive protein database in which the reviewed section contains expert-annotated information from the literature, including PTMs [4]
2. PRO provides a structured hierarchical representation of protein entities and protein complexes [9]. It consists of three sub-ontologies: ProEvo, for evolutionary relationships among

proteins; ProForm, for relationships among proteoforms (*see Note 3*), including PTM forms of a protein; and ProComp, for protein complexes. Two features of PRO are particularly relevant to iPTMnet. First, orthologous proteins can be easily identified using PRO. In the PRO hierarchy, species-specific orthologs of a protein are connected by an *is_a* relation (parent-child relationship) to a species-independent term. Second, PRO terms can be defined for individual PTM proteoforms, including forms with multiple PTMs, and these terms can be associated with proteoform-specific functional information. For example, the mouse apoptosis regulatory protein BAD has three phosphorylated proteoforms in PRO involving different combinations of three sites: S112, S136, and S155. PR:000026136 is phosphorylated on S136; PR:000026167 is phosphorylated on S112 and S136; and PR:000026133 is phosphorylated on S112, S136, and S155. The S112/S136 doubly phosphorylated form (PR:000026167) exhibits a phosphorylation-dependent decreased interaction with 14-3-3 proteins; this information is included in the term's annotation. The relationship among PTM proteoforms of a protein can be easily seen in the PRO hierarchical structure.

3. Proteoform refers to any of the protein products of a single gene, including those that arise by genetic mutation, alternative splicing, and post-translational modification [20].
4. The data and screenshots presented in this article were collected from iPTMnet v3.1 in May 2016.
5. The Substrate column of the PTM enzyme page displays the UniProtKB identifier followed by the gene name in parentheses. If no gene name is given in the UniProtKB record, then only the UniProtKB identifier will be displayed. For example, no gene name is displayed for the MPK6 substrate Q9SJF3 (Fig. 5a) because the UniProtKB record does not provide this information. On the UniProtKB entry page, the gene is referred to by its *Arabidopsis* ordered locus identifier, At5g67300.
6. If the PTM enzyme is listed as MPK6 (with no proteoform information), it does not imply that the MPK6 is unphosphorylated, just that its phosphorylation status was not assayed and/or reported. It is possible (and even likely given what we know about MAPK signaling pathways and MAPK activity) that MPK6 is phosphorylated on its activating sites in all cases.

Acknowledgments

This work was funded by grants from the National Institutes of Health (U01GM120953 and P20GM103446).

References

- Minguez P, Letunic I, Parca L, Bork P (2013) PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic Acids Res* 41(Database issue):D306–D311. doi:10.1093/nar/gks1230
- Patterson H, Nibbs R, McInnes I, Siebert S (2014) Protein kinase inhibitors in the treatment of inflammatory and autoimmune diseases. *Clin Exp Immunol* 176(1):1–10. doi:10.1111/cei.12248
- Zhou N, Xu W, Zhang Y (2015) Histone deacetylase inhibitors merged with protein tyrosine kinase inhibitors. *Drug Discov Ther* 9(3):147–155. doi:10.5582/ddt.2015.01001
- UniProt C (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43(Database issue):D204–D212. doi:10.1093/nar/gku989
- Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 43(Database issue):D512–D520. doi:10.1093/nar/gku1267
- Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res* 39(Database issue):D261–D267. doi:10.1093/nar/gkq1104
- Torii M, Arighi CN, Gang L, Qinghua W, Wu CH, Vijay-Shanker K (2015) RLIMS-P 2.0: a generalizable rule-based information extraction system for literature mining of protein phosphorylation information. *IEEE/ACM Trans Comput Biol Bioinform* 12(1):17–29. doi:10.1109/TCBB.2014.2372765
- Tudor CO, Ross KE, Li G, Vijay-Shanker K, Wu CH, Arighi CN (2015) Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. *Database (Oxford)* 2015. doi:10.1093/database/bav020
- Natale DA, Arighi CN, Blake JA, Bult CJ, Christie KR, Cowart J, D'Eustachio P, Diehl AD, Drabkin HJ, Helfer O, Huang H, Masci AM, Ren J, Roberts NV, Ross K, Ruttenberg A, Shamovsky V, Smith B, Yerramalla MS, Zhang J, AlJanahi A, Celen I, Gan C, Lv M, Schuster-Lezell E, Wu CH (2014) Protein ontology: a controlled structured network of protein entities. *Nucleic Acids Res* 42(Database issue):D415–D421. doi:10.1093/nar/gkt1173
- Colcombet J, Hirt H (2008) Arabidopsis MAPKs: a complex signalling network involved in multiple biological processes. *Biochem J* 413(2):217–226. doi:10.1042/BJ20080625
- Nuhse TS, Peck SC, Hirt H, Boller T (2000) Microbial elicitors induce activation and dual phosphorylation of the Arabidopsis thaliana MAPK 6. *J Biol Chem* 275(11):7521–7526
- Valenta T, Hausmann G, Basler K (2012) The many faces and functions of beta-catenin. *EMBO J* 31(12):2714–2736. doi:10.1038/emboj.2012.150
- Bauer A, Lickert H, Kemler R, Stappert J (1998) Modification of the E-cadherin-catenin complex in mitotic Madin-Darby canine kidney epithelial cells. *J Biol Chem* 273(43):28314–28321
- Breitschopf K, Haendeler J, Malchow P, Zeiher AM, Dimmeler S (2000) Posttranslational modification of Bcl-2 facilitates its proteasome-dependent degradation: molecular characterization of the involved signaling pathway. *Mol Cell Biol* 20(5):1886–1896
- Furukawa Y, Iwase S, Kikuchi J, Terui Y, Nakamura M, Yamada H, Kano Y, Matsuda M (2000) Phosphorylation of Bcl-2 protein by CDC2 kinase during G2/M phases and its role in cell cycle regulation. *J Biol Chem* 275(28):21661–21667. doi:10.1074/jbc.M906893199
- Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human protein reference database—2009 update. *Nucleic Acids Res* 37(Database issue):D767–D772. doi:10.1093/nar/gkn892
- Zulawski M, Braginets R, Schulze WX (2013) PhosPhAt goes kinases—searchable protein kinase target information in the plant phosphorylation site database PhosPhAt. *Nucleic Acids Res* 41(Database issue):D1176–D1184. doi:10.1093/nar/gks1081
- Yao Q, Ge H, Wu S, Zhang N, Chen W, Xu C, Gao J, Thelen JJ, Xu D (2014) P(3)DB 3.0: from plant phosphorylation sites to protein networks. *Nucleic Acids Res* 42(Database issue):D1206–D1213. doi:10.1093/nar/gkt1135

19. Sadowski I, Breitkreutz BJ, Stark C, Su TC, Dahabieh M, Raithatha S, Bernhard W, Oughtred R, Dolinski K, Barreto K, Tyers M (2013) The PhosphoGRID *Saccharomyces cerevisiae* protein phosphorylation site database: version 2.0 update. Database (Oxford) 2013:bat026. doi:[10.1093/database/bat026](https://doi.org/10.1093/database/bat026)
20. Smith LM, Kelleher NL, Consortium for Top Down P (2013) Proteoform: a single term describing protein complexity. Nat Methods 10(3):186–187. doi:[10.1038/nmeth.2369](https://doi.org/10.1038/nmeth.2369)

Part IV

Proteomic Bioinformatics

Chapter 17

Protein Identification from Tandem Mass Spectra by Database Searching

Nathan J. Edwards

Abstract

Protein identification from tandem mass spectra is one of the most versatile and widely used proteomics workflows, able to identify proteins, characterize post-translational modifications, and provide semiquantitative measurements of relative protein abundance. This manuscript describes the concepts, prerequisites, and methods required to analyze a tandem mass spectrometry dataset in order to identify its proteins, by using a tandem mass spectrometry search engine to search protein sequence databases. The discussion includes instructions for extraction, preparation, and formatting of spectral datafiles, selection of appropriate search parameter settings, and basic interpretation of the results.

Key words Protein identification, MS/MS spectra, Protein sequence databases, Peptide identification, Search engine

1 Introduction

The identification of proteins by tandem mass spectrometry is one of the most versatile and widely used techniques in mass spectrometry-based proteomics. It can be applied to purified protein samples containing just a few proteins, or to complex samples containing many proteins, such as cell lines or clinical tissue samples. Tandem mass spectrometry can be applied without prior knowledge of the protein content of the analytical sample and can readily identify proteins, characterize post-translational modifications, and provide semiquantitative measurements of relative protein abundance. Furthermore, the acquisition of tandem mass spectra by modern mass spectrometers is highly automated, making it possible to conduct high-throughput, comprehensive analyses of complex protein mixtures, generating tens to hundreds of thousands of tandem mass spectra per sample and facilitating the characterization of complex proteomes.

The most common application of tandem mass spectrometry in proteomics workflows seeks to identify the unknown proteins of

an analytical sample. In a workflow called shotgun proteomics, by analogy with whole genome shotgun sequencing, or bottom-up proteomics, to contrast with top-down, intact protein analysis, the sample's proteins are solubilized and subjected to proteolysis to produce short peptides, 10–20 amino acids in length, for analysis by (tandem) mass spectrometry. The resulting peptide mixture is separated according to the peptides' physical and chemical properties using liquid chromatography and analyzed by mass spectrometry as they elute off the column. As peptides with similar physical and chemical properties elute from the column, the mass spectrometer collects survey scans and selects abundant peptide ions for fragmentation and analysis by tandem mass spectrometry, resulting in MS/MS or tandem mass spectra. Mass spectrometers can collect as many as 50 tandem mass spectra after each survey scan, with each cycle of one survey scan and many tandem mass spectra taking a few seconds. Over the course of a 2–4 h chromatography gradient, this shotgun/bottom-up proteomics workflow can collect tens to hundreds of thousands of MS/MS spectra, each representing a fragmented peptide from the sample's proteins. The automated acquisition of tandem mass spectra in conjunction with liquid chromatography is often shortened to LC-MS/MS.

Computer software called tandem mass spectrometry search engines analyze shotgun/bottom-up proteomics spectra datasets to identify the proteins in the sample. These search engines match the tandem mass spectra with peptide sequences from protein sequence databases and use the confidently matched peptides to infer the protein content of the sample. This manuscript describes the concepts, prerequisites, and methods required to analyze shotgun/bottom-up proteomics datasets using tandem mass spectrometry search engines. The discussion includes instructions for extraction, preparation, and formatting of spectral datafiles; selection of appropriate search parameter settings; and interpretation of the results. For more background on these various techniques, we refer the reader to one of the many excellent reviews [1–7].

This manuscript will focus on the analysis of a typical shotgun/bottom-up proteomics dataset acquired from a complex protein mixture for the purposes of protein identification. The various experimental technologies of mass spectrometry-based proteomics are constantly changing, with improved instrument resolution, fragmentation analysis of larger peptides and proteins, new ionization and fragmentation technologies, and novel separation techniques and digest reagents. We will not explicitly discuss these new technologies, though we expect that this manuscript will provide a foundation for understanding how software tools might be used for this data, too.

This manuscript will not directly address the analysis of datasets from proteomics quantitation workflows, except where these impact the settings used for protein identification from tandem mass spectra. This manuscript will also not attempt to describe, in

explicit terms, instructions for any specific search engine, but will document the concepts and principles behind each of the typical parameters so that appropriate choices can be made for any search engine. Finally, we note that we do not address a variety of other techniques and software tools for protein identification from tandem mass spectra, notably the *de novo* [8–11] and hybrid sequence tag [12–14] approaches.

2 Concepts

In order to streamline the description of the methods to follow, we introduce some important mass spectrometry concepts and terminology that will be used throughout the manuscript.

2.1 Ionization

Mass spectrometry is an analytical technique that measures the mass of molecules and atoms [15]. The molecules to be analyzed are transformed into charged, gas-phase *ions* which can be manipulated and detected by the mass spectrometer. The *ionization* of peptides for mass spectrometry is typically carried out using one of two technologies: *electrospray ionization (ESI)* or *matrix-assisted laser desorption/ionization (MALDI)*. Peptides are given charge during ionization in positive ion mode by *protonation*, the addition of one or more protons, with the peptides observed in shotgun/bottom-up proteomics experiments typically gaining one proton when subject to MALDI ionization, and one to five protons when subject to ESI ionization.

2.2 Mass-to-Charge Ratio

The mass spectrometer's mass analyzer uses electrical, magnetic, and RF fields to separate the gas-phase ions in time or space before they are detected and counted. These fields manipulate the ions based on their *mass-to-charge ratio (m/z value)*, rather than their mass; therefore, the number of attached protons must be determined before the mass of an ion can be inferred. The number of protons attached to a peptide or fragment ion is called its *charge state*. The *x-axis* of a mass spectrum, such as the example in Fig. 1, records observed ions' *m/z* values in atomic mass units, the approximate mass of a hydrogen atom, or equivalently, that of a proton or a neutron. The mass spectrometry community generally refers to atomic mass units as *Daltons (Da)*. We point out that the protonation of a peptide affects not only its *m/z* value but also its mass, increasing the mass by 1.0078 Da. Conversion between the molecular weight (*m*) and the mass-to-charge ratio (*m/z*) of a positively charged ion in charge-state *z*+ can be accomplished using the formula:

$$m/z = \frac{(m + zm_H)}{z} \quad (1)$$

where $m_H = 1.0078$ is the mass of a proton/hydrogen, in Daltons, to five significant figures.

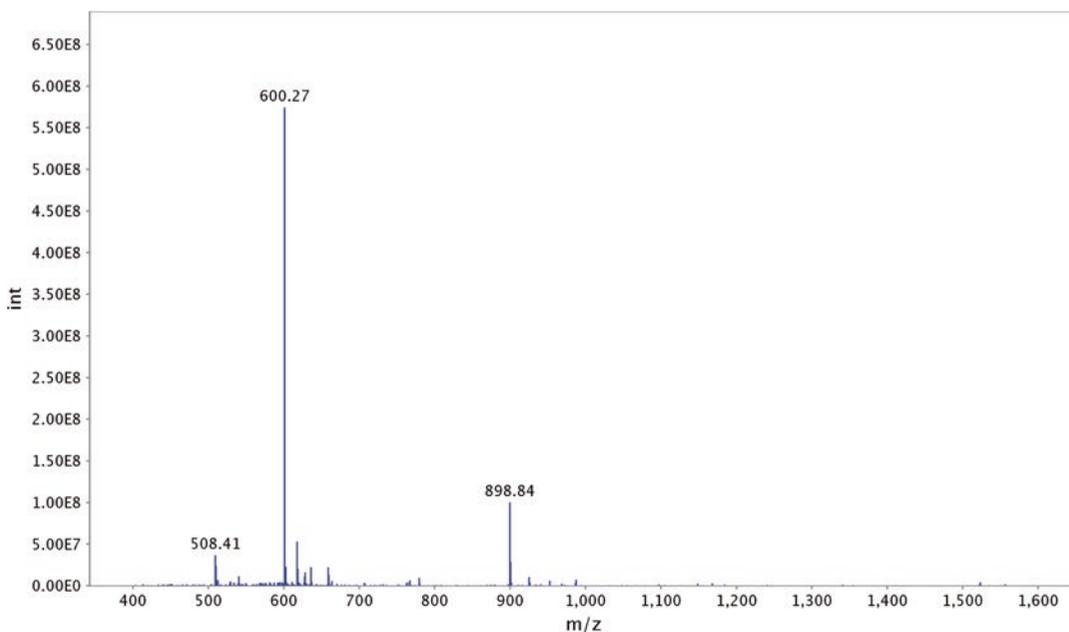


Fig. 1 Survey scan (#984) from spectra file raffflow37 from Peptide Atlas dataset raffflow (<http://www.peptide-atlas.org/repository/>; Accession Number PAe000004)

2.3 Ion Counts and Resolution

Mass spectra are essentially histograms of *ion counts*, with the y -axis of a mass spectrum representing the number of ions at a particular m/z value in arbitrary units, as shown in Fig. 1. The ion counts for an individual molecular ion are observed at many similar m/z values, generating a characteristic peak shape. A spectrum that samples each ions' peaks at many similar m/z values is called a *profile spectrum*. The ability of a mass spectrometer to distinguish different ions with distinct m/z values, called *resolution*, depends on their true m/z values and the width of the peak shape observed in the profile mass spectrum. Figure 2 shows the peak shapes for peptide ions of AACLLPKLDEL RDEGK, demonstrating how the peaks of individual ions may overlap, or convolve, as they get close together. Low-resolution instruments may ultimately be unable to measure the m/z values of individual ions if they get too close, an effect which has important consequences for determining peptide ion charge states (*see* Subheading 2.5) and therefore mass.

2.4 Peak Detection

After acquisition, profile spectra are analyzed using *peak detection* algorithms, to obtain *centroided spectra*, or *peak lists*, in which each ion's peak shape is integrated and summarized by m/z value and integrated intensity. After centroiding, the peaks of Fig. 2 become impulses, without shape, and depending on the peak detection algorithm and instrument resolution, may represent the underlying, convolved peak shapes of peptide ions accurately, or in some cases, poorly. Figure 3 shows the centroided spectrum result of a simple apex-based peak detection algorithm applied to each of the

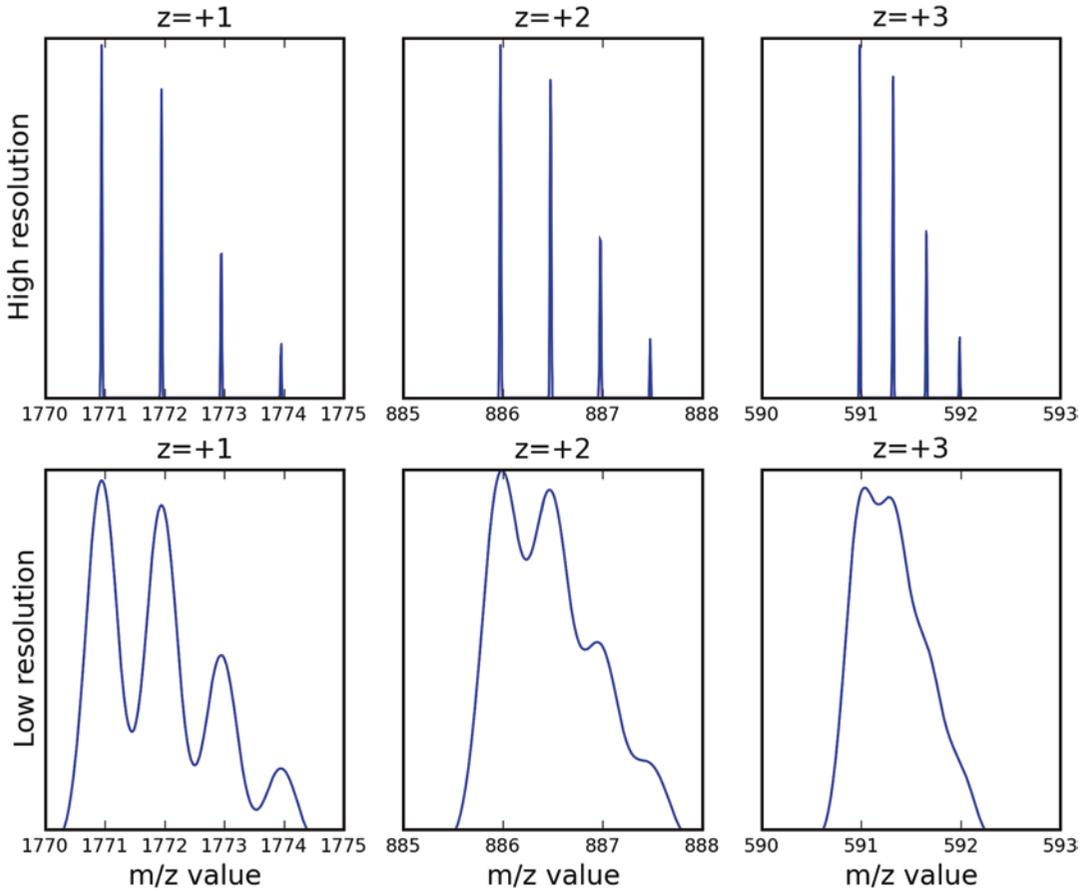


Fig. 2 Schematic isotope clusters for peptide AACLLPKLDEL RDEGK (molecular weight 1769.93 Da), in charge states 1+, 2+, and 3+, as might be observed in profile spectra from high (*top row*) and low (*bottom row*)-resolution instruments

profile spectra of Fig. 2. While the low-resolution peak shapes for charge states 2+ and 3+ in this figure represent the convoluted shapes of four distinct molecular ions, most peak detection algorithms will output just two or three peaks in this region. Unsurprisingly, the result of peak detection on the high-resolution profile spectra captures the peptides' ions well.

2.5 Isotope Clusters and Charge State

Peptides, as naturally occurring organic molecules, contain elements such as carbon which sometimes incorporate one or more extra neutrons in the nucleus. Approximately 1 % of naturally occurring carbon is observed as the ^{13}C isotope rather than the more common ^{12}C isotope. When one ^{13}C isotope is present in a peptide, its molecular mass increases by 1.0025 Da (to 5 s.f). In any proteomics sample, the masses observed for the many copies of a particular peptide are probabilistically distributed among those with no ^{13}C isotopes, exactly one ^{13}C isotope, exactly two ^{13}C

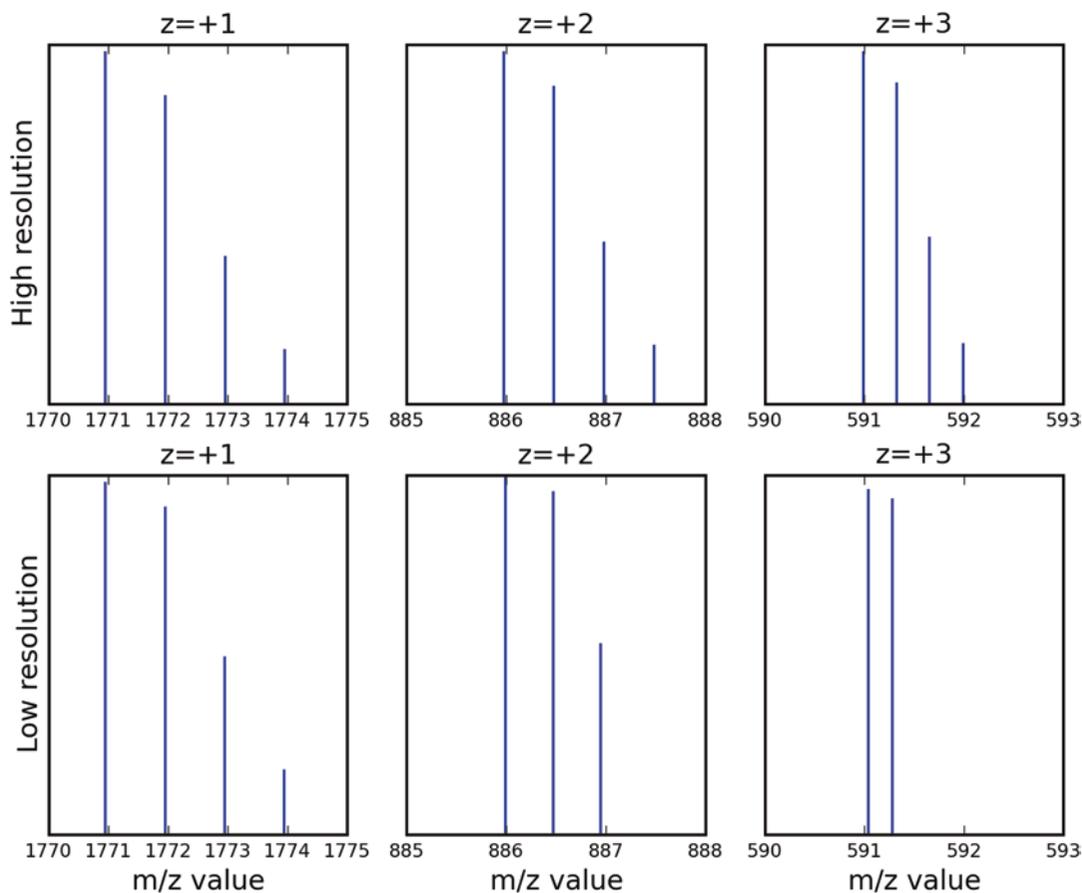


Fig. 3 Schematic isotope clusters for peptide AACLLPKLDEL RDEGK (molecular weight 1769.93 Da) after peak detection, in charge state 1+, 2+, and 3+, as might be observed in centroided spectra from high- (*top row*) and low (*bottom row*)-resolution instruments

isotopes, and so on. As the mass of the peptide increases, the probability of incorporating ^{13}C isotopes increases too, so the relative intensities of each of the peptide's *isotopic peaks* changes with mass. Figure 2 shows an *isotope cluster* for the peptide AACLLPKLDEL RDEGK (molecular weight 1769.93 Da) in charge states 1+, 2+, and 3+, as might be observed in a low-resolution and high-resolution mass spectrometer. Notice that the peaks of the 1+, 2+, and 3+ charge state isotope clusters are separated by 1, 1/2, and 1/3 Daltons, making it possible to infer the charge state of the ion. Depending on the instrument's resolution, however, we may not be able to reliably determine the m/z value spacing of the isotope cluster peaks, particularly after peak detection, in order to establish the charge state of peptide ions.

2.6 Average vs Monoisotopic Mass

Depending on the resolution of the instrument and the mass of the molecules being analyzed, the result of peak detection may not reflect a specific isotope of a peptide ion, but may instead represent the centroid and peak integration of the convolution of all of the peptide ions' isotopes' peak shapes. In this case, the reported m/z value for the peak represents a weighted average of the individual isotope cluster peaks, and the intensity represents the peak integration of the entire cluster. This m/z value represents the *average mass* of the peptide ion. With sufficient resolution or small enough masses, a centroided peak representing the *monoisotopic mass* of the peptide ion can be reliably determined. The monoisotopic mass is the mass of the peptide ion calculated using only the most abundant isotopic form of each element, which is the leftmost peak of the isotope cluster for peptides in typical shotgun/bottom-up proteomics experiments. The calculated monoisotopic mass of peptide AACLLPKLDEL RDEGK of Fig. 2 in charge state 1+ is 1770.94 Da, while the average mass is 1771.93 Da.

2.7 Peptide Fragmentation Spectra

Peptide tandem mass spectra measure the m/z values of fragment ions formed by collisionally induced dissociation (CID), in which precursor ions break apart due to collisions with pressurized, inert gas molecules. Precursor ions, selected by their m/z value, represent many copies of an ionized peptide. When the peptide ions break apart in CID, their protons are retained by one or more of the fragments, and the charged fragment ions measured by the mass analyzer and detector, forming the tandem mass spectrum. Peptides tend to fragment along the peptide's amino acid backbone, revealing the peptide's primary structure, its amino acid sequence. When the N-terminus (left end) fragment retains a proton, the fragment ions are named using the initial letters of the alphabet and the number of amino acids in the fragment. Similarly, when the C-terminus (right end) fragment retains a proton, its fragment ions are named using the last letters of the alphabet. The most common fragment ions formed by low-energy CID are the *b-ions* and *y-ions*, though *a-ions* are also observed sometimes. Figure 4 shows a peptide fragmentation spectrum of the precursor ion at m/z value 898.84 of Fig. 1 representing the peptide LQGSATAAEA QVGHQTAR in charge state 2+. Observed *b-* and *y-*ions are marked in the spectrum and on the peptide sequence logo in the top-left corner.

The molecular weight of N-terminus peptide fragment ions is computed using the formula

$$t_i = m_N + \sum_{k=1}^i m_{aa_k} \quad \text{for } i = 1, \dots, n-1 \quad (2)$$

where ion-type $t \in \{a, b, c\}$; m_{aa_k} is the residual mass of the k th amino acid of the peptide, $k = 1, \dots, n$; the N-terminal offset

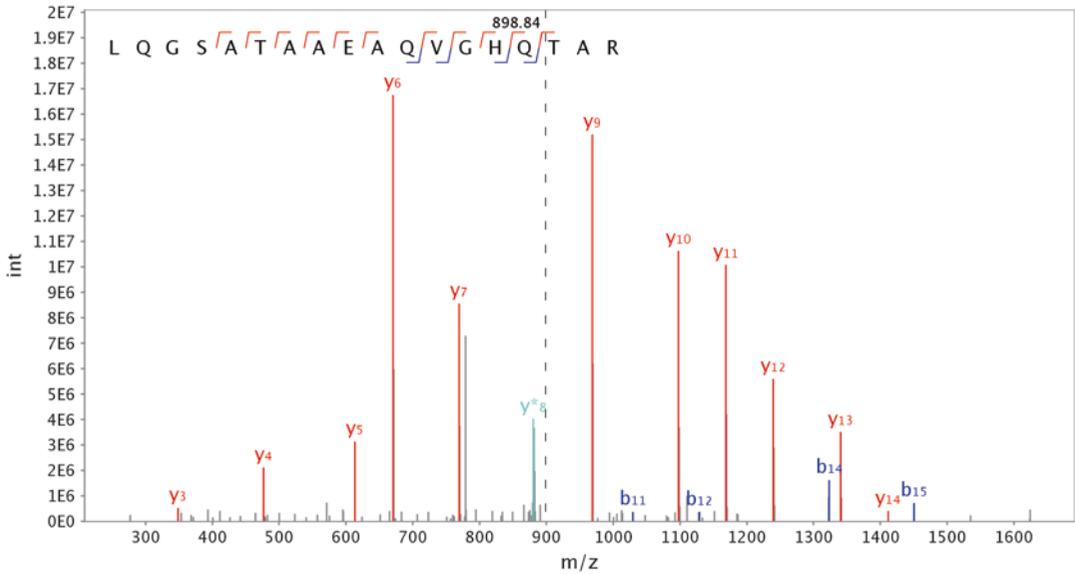


Fig. 4 Tandem mass spectrum (scan #985) of peptide LQGSATAAEQVGHQATAR, from the charge state 2+ precursor ion with m/z value 898.84 in scan #984 (Fig. 1), from spectra file raffflow37 of the Peptide Atlas dataset raffflow

$m_N=0.0$; and the ion-type offsets $m_a = -27.9994$; $m_b=0.0$; and $m_c = 17.0265$. The m/z values are determined using (1).

The molecular weight of C-terminus peptide fragmentation ions are computed using the formula

$$t_i = m_C + \sum_{k=1}^i m_t \quad \text{for } i = 1, \dots, n-1 \quad (3)$$

where $t \in \{x, y, z\}$; the C-terminal offset $m_C = 18.0106$, and the ion-type offsets $m_x = 25.9793$; $m_y = 0.0$; and $m_z = -16.0187$. The m/z values are determined using (1).

3 Materials

3.1 Tandem Mass Spectrometry Search Engine

There are many open-source and commercial tandem mass spectrometry search engines available, although only a few have become widely used. Commercial search engines Mascot [16], from company Matrix Science, and SEQUEST [17], available from Thermo Fisher, are very popular. More recently, a number of free, open-source search engines have found significant adoption, particularly X!Tandem [18] and MSGF+ [19].

For small-scale or *ad hoc* analyses, Mascot and X!Tandem are available for free on the web, at the websites shown in Table 1.

Table 1
Free web-based interfaces to popular tandem mass spectrometry search engines, suitable for small-scale or ad hoc analyses

Search engine	URL
Mascot	http://matrixscience.com
X!Tandem	http://thegpm.org

Table 2
Web-based interfaces for large-scale analyses of tandem mass spectrometry data

Search tool	URL
PepArML	http://grg.tn/peparml
TPP	http://tools.proteomecenter.org/twa
Chorus	http://chorusproject.org
ProteoSAFE	http://massive.ucsd.edu

Each of these web-based tools allows the user to upload a spectral datafile, select search parameters, execute the search, and browse and interpret the results. For large-scale analyses or where greater configuration flexibility is required, users can either download, and perhaps compile, the free search engines and configure searches to run on the available local compute resources or use services available in the cloud under a variety of cost models. Examples of these services include the PepArML [20] search engine, the Trans-Proteomic Pipeline in the cloud [21], Chorus, and ProteoSAFE. See Table 2.

3.2 Protein Sequence Database

The protein sequence database provides peptide sequences to be matched against the tandem mass spectra by the search engine. As such, the selected protein sequence database will have a significant impact on the sensitivity, specificity, and speed of the search. Since peptides whose sequence is missing from the protein sequence database cannot be matched with any spectrum, poorly chosen sequence databases will result in spectra going unidentified and their peptides being unobserved. However, larger, more inclusive protein sequence databases take longer to search and may result in more false-positive identifications and reduced statistical significance.

Locally installed search engines generally expect FASTA format protein sequence databases, which can be readily downloaded from a variety of websites. Installation of protein sequence data-

bases for locally installed search engines may require special configuration and preprocessing of the protein sequence database file, but the analysis flexibility gained is significant. The local installation of specific protein sequence databases is one of the primary reasons to install and run a search engine in-house, as the sequence databases provided by the free web-based search engines are often quite limited.

When available, organism-specific sequence databases eliminate false positives from closely related species, but can leave peptides from common contaminants unidentified. For this reason, keratins, trypsin, and other artifactual protein sequences are sometimes added to organism-specific sequence databases, even though they do not inform the biology of the sample.

Where the source of the proteins is a single, well-characterized model organism, UniProt [22] reference proteomes, downloaded with isoforms, are a good choice. If the origin of the sample is unknown, or known to be a mixture of organisms, then the Swiss-Prot section of UniProtKB [22] is a good choice. UniProt also provides tools for selecting and downloading sub-proteomes constrained by protein feature or annotation. NCBI's RefSeq is another good source of protein sequences and is available in various taxonomic divisions. Organism-specific RefSeq sequences can be found in the genomes section of the NCBI FTP site.

The use of NCBI's nr and similar computationally merged protein sequence databases or protein sequences from poorly annotated genomes is not recommended as redundant peptide sequences, and poor protein naming can significantly complicate the interpretation of the results. In some cases, searching ESTs and genome sequences may be appropriate [23], but considerable post-search analysis must be carried out to make up for the absence of reliable protein annotation.

3.3 Instrument Characteristics

The technical characteristics of the mass spectrometer and its configuration for a particular experiment have a large impact on the tandem mass spectra acquired. As such, these details inform the selection of appropriate search parameters. Some parameters do not change on an experiment to experiment basis, but others do—in each case, the informatics analyst may need to consult with the mass spectrometrists to determine appropriate settings.

It is crucial to establish whether the ionization (*see* Subheading 2.1) technology used is MALDI or ESI, as MALDI ionization primarily generates charge state 1+ peptide ions. Electrospray (ESI) instruments usually generate peptide ions in charge states 1+, 2+, and 3+, although 4+ and 5+ charge states are sometimes observed for larger, more basic peptides. Charge state 1+ precursors can be assumed for MALDI tandem mass spectra, significantly simplifying their analysis. In a pinch, the absence of 2+ and 3+ precursors and a preponderance of 1+ precursors can be taken to indicate MALDI

MS/MS spectra, while precursors with charge states greater than 1+ indicate ESI spectra.

Appropriate mass-accuracy parameters for precursor and fragment ion m/z values must reflect realistic performance characteristics of the mass spectrometer. These parameters determine when an experimental m/z value is considered to *match* the peptide or fragment mass computed in silico. Fragment ion matches are the foundation of scoring and evaluation of peptide identifications. If the fragment mass-accuracy parameter is set too tightly, many valid fragment matches will be missed, reducing true positive peptides' scores. If the fragment mass-accuracy parameter is set too loosely, spurious fragment matches will be observed, inflating all peptides' scores and increasing the number of false-positive identifications.

Precursor mass-accuracy parameters should also be chosen carefully to reflect the resolution of the survey scan. The precursor mass-accuracy parameter is the primary criteria used by the search engines to determine whether or not a peptide sequence will be considered for scoring against a particular spectrum. If set too tight, many valid peptide identifications will be missed. If set too loose, many incorrect peptide sequences will be scored, resulting in slower search times and increased potential for false-positive identifications. Sometimes, the instrument's real-time algorithm for picking precursor ions from the survey scan will select isotope cluster peaks (*see* Subheading 2.5) other than the monoisotopic peak—in this case, the search engine will require a precursor match tolerance of at least 1 Da to match the experimental m/z value with the monoisotopic mass computed from the peptide sequence. As the detrimental effects of loose precursor mass-accuracy parameters are generally mild, a precursor mass tolerance of 2.5 Da is often used to ensure these peptide ions are identified.

Note that for some instruments, particularly the LTQ-Orbitrap (Thermo Fisher), the mass analyzer and the resolution used for survey scans and tandem mass spectra in a particular experiment are configured by the mass spectrometrist at acquisition time—for these instruments, the instrument model is not sufficient to establish mass-accuracy parameters.

For older instruments, it is important to understand whether the reported m/z values represent average masses or not (*see* Subheading 2.6). Current instruments are generally able to resolve individual isotope cluster peaks for the relatively low charge states in typical LC-MS/MS data, so usually experimental m/z values should be matched against monoisotopic masses rather than average masses.

3.4 Tandem Mass Spectra Datafile

The handling and pre-processing of tandem mass spectra datafiles can, unsurprisingly, have significant impact on the peptide identification results. Mass spectrometers generally store spectral data in binary vendor (and instrument)-specific file formats that can only

be read by software provided by the instrument vendor. The mass spectrometer's "raw" spectra files must usually be processed and exported to some nonproprietary, open format for analysis by tandem mass spectrometry search engines. Typically, the vendor software will provide some facility for tandem mass spectra export, and a number of open-source tools that use the vendor libraries are available as part of the Trans-Proteomic Pipeline¹ (TPP) [24] or the ProteoWizard² [25] projects.

Peak detection or centroiding (*see* Subheading 2.4) *must* be carried out if the raw spectra files contain profile spectra, as the tandem mass spectrometry search engines require peak lists from centroided spectra. Some tools provide additional spectral processing facilities, which can improve the quality of the spectra to be analyzed. Common spectral processing options include intensity thresholding, deisotoping, precursor charge state determination or enumeration, and spectrum merging or averaging.

Intensity thresholding removes peaks less than a specific relative intensity, which can reduce spectrum size and eliminate spurious fragment matches. Deisotoping finds isotope clusters and eliminates the non-monoisotopic peaks, reducing spurious fragment matches. Both of these processing steps have the potential to do more harm than good, removing valid fragment ions from the spectrum. Precursor charge states may be determined by examining isotope clusters (*see* Subheading 2.5) in the survey scans, which are often not exported for the tandem mass spectrometry search engine. Alternatively, multiple copies of each tandem mass spectrum may be enumerated, each with a different declared charge state. Spectral merging and averaging can be carried out when the same precursor m/z value is selected for fragmentation multiple times in an LC-MS/MS experiment. Merging or averaging these tandem mass spectra boosts the intensity of common peaks and reduces the intensity of noise and can have a significant impact on the peptide identification results.

The open-source tools tend not to provide deisotoping, precursor charge state determination, or spectral averaging facilities, and where the vendor libraries do not provide peak detection routines, the open-source tools implement relatively crude centroiding algorithms. Nevertheless, until quite recently, vendors did not provide tools for exporting the spectra in convenient open formats, and open-source tools were the only option. Due to the dependence on vendor software and libraries, raw spectra conversion must generally be carried out on the Windows platform.

Commonly used open file formats for tandem mass spectra are indicated by their file extension: .dta, representing a single tandem mass spectrum, with the filename encoding scan number and

¹ See <http://tools.proteomecenter.org/wiki/index.php?title=Formats:mzXML>.

² See <http://proteowizard.sourceforge.net>.

charge state information; .mgf (Mascot generic format), representing many tandem mass spectra in a simple, text-based format; and .mzXML and .mzML, representing many tandem mass spectra and their metadata using XML. Of the XML formats, mzXML appeared first and has been widely adopted. mzML represents the current HUPO Proteomics Standards Initiative XML standard for tandem mass spectrometry data.

A significant issue with some of these formats is the loss of important metadata associated with each tandem mass spectrum, particularly the original raw datafile *scan numbers* and LC retention times. Using mzML is the best way to avoid losing this metadata. The MGF format is the most widely supported of the generic, simple, text-based tandem mass spectrometry formats.

3.5 Sample Preparation

Information about the manipulation and handling of the protein sample prior to analysis by mass spectrometry is the final prerequisite for a successful peptide identification analysis. First, the proteolytic enzyme used to cleave proteins into peptides must be established. Trypsin, which cuts at Arg (R) and Lys (K) unless followed by Pro (P), is the most common. Second, Cys (C) residues are typically chemically modified, deliberately, to ensure they have a known, predictable mass. Iodoacetamide is the most commonly used reagent for this purpose, subjecting the Cys residues to carbamidomethylation and increasing their mass by 57.0215 Da. Other deliberate chemical labeling of specific residues and peptide or protein termini should also be noted, since these too change the expected masses of peptides. In particular, many proteomics quantitation workflows use stable isotope labels, differential chemical modifications on specific amino acids, or peptide termini, and these must be considered in setting appropriate peptide identification search parameters.

Where the sample represents a single protein (extracted from a gel band or spot) or is derived from a specific species or a specific biological context, this should also be noted, as this can be useful for selection of protein sequence databases and interpreting the results.

4 Methods

4.1 Prepare the Spectra Datafile

The vendor and open-source tools available for processing and exporting the tandem mass spectra may not output the spectra in a format supported by your choice of search engine. Once in *some* open format, however, any number of tools are available for reformatting the spectra. The TPP and ProteoWizard projects (*see* Subheading 3.4) provide a number of programs for converting between a variety of XML formats and the more basic formats, such as dta and mgf.

4.2 Specify Search Parameters

Many of the search parameters required by tandem mass spectrometry search engines impact the classic balance between search time and the potential to miss valid peptide identifications. The informatician must ensure that the parameters chosen do not exclude too many valid peptide identifications while keeping the search time reasonable. For these parameters, there is always the question of whether or not a more thorough search will yield sufficient additional identifications (spectra, peptides, or proteins) to justify the additional search time. While the search time consequences must always be paid, the benefit is generally impossible to quantify before searching, unless some kind of additional information is available. We make these trade-offs explicit, where they are relevant, in the following steps.

4.2.1 Protein Sequence Database

The selection of the protein sequence database to search represents the most significant trade-off between search time and the potential to miss peptide identifications due to the absence of true-positive peptide sequences from the protein sequence database. Larger, more inclusive protein sequence databases will take longer to search, but may identify more peptides. Smaller, more selective protein sequence databases will take less time to search, but important or unexpected peptides may be missed. Where the source of the proteins is a single, well-characterized model organism, the UniProt reference proteomes are a good choice. If the origin of the sample is unknown, or known to be a mixture of organisms, then the Swiss-Prot section of UniProtKB is a good choice. For web-based search engines, some interesting options may be available, depending on the site, but specific exotic options, such as proteins from a specific bacterial genome, may not. *See* Subheading 3.2 for a discussion of the sequence database options for locally installed search engines.

4.2.2 Instrument Parameters

Having established the instrument characteristics as a prerequisite to the search in Subheading 3.3, all that remains is to map these characteristics to the parameters required by the search engine. Average or monoisotopic masses should be specified as appropriate (*see* Subheadings 2.6 and 3.3). Mass tolerance parameters are generally specified in Da (Daltons) or ppm (parts per million). The ppm units are used when the mass tolerance is proportional to the measured mass, while Da is used when the mass tolerance is constant with the measured mass. Low-resolution tandem mass spectra may require a fragment mass match tolerance setting as large as 0.6 Da, while for higher-resolution fragmentation spectra, a setting of 0.1 Da is appropriate.

Some search engines will require the name of the instrument or a vendor neutral abbreviation of its ionization and mass analyzer technologies to be specified, since these can significantly affect the characteristics of peptide fragmentation.

4.2.3 Mass Modification Parameters

While the residual, unmodified mass of amino acids is well established, there is no guarantee that the particular peptide ion observed in the mass spectrometer contains only unmodified amino acids. Some residues, particularly Cys, are chemically modified deliberately (*see* Subheading 3.5) as part of the sample preparation. In this case, the mass modification is called *fixed* and is applied to every Cys residue of every peptide, before scoring. There is no running time penalty for fixed mass modifications. Incorrectly setting or failing to set the Cys fixed modification, however, will render the spectra of Cys containing peptides unidentifiable. The carbamidomethylation of Cys is the most common of such modification, and if in doubt, a +57.0215 fixed mass modification, representing carbamidomethylation of Cys by iodoacetamide, should be used.

So-called variable modifications indicate potential adducts on specific residues. If the oxidized Met variable modification is selected, then every Met residue in the sequence database will be considered with both the nominal mass of 131.0405 Da and the mass with an additional 15.9949 Da (147.0354 Da). Variable modifications can be specified for biological mass modifications, such as for phosphorylation, or common artifactual mass modifications, such as oxidation, on specific residues. The search time increases exponentially in the number of variable modifications, so they should only be used when they are expected.

4.2.4 Proteolytic Enzyme Parameters

The proteolytic enzyme setting should match the sample preparation conditions established in Subheading 3.5.

Even when the specific proteolysis enzyme and its cleavage motif is known, there is no guarantee that it will cleave at every site the motif appears or that it will leave non-motif positions alone. As such, the samples' peptides may ultimately have none, one, or both termini consistent with the enzyme and may contain internal motif sites representing a missed cleavage opportunity. By default, search engines will consider only those peptide sequences with both N- and C-terminus consistent with the selected proteolytic enzyme. A semi-specific (or semi-tryptic) search will consider peptide sequences with at least one of the N- or C-termini consistent with the proteolytic enzyme (trypsin). A nonspecific search will consider all peptide sequences, regardless of the N- or C-termini sequence. The selection of semi-specific proteolysis will typically increase search times by a factor of 20–30, but can sometimes increase the number of identified peptides substantially. A nonspecific search is usually only applied in special cases.

The maximum number of internal motif sites a peptide may contain is controlled by a parameter called missed cleavages, which is typically set to a small number, such as 1 or 2.

4.2.5 Precursor Mass Tolerance Parameters

As outlined in Subheading 3.3, appropriate settings for the precursor mass tolerance parameters should be set to ensure that peptide sequences match the precursor mass of their spectra.

Older instruments with poor precursor mass accuracy and significant potential for selection of non-monoisotopic isotope cluster peaks as precursor ions have led to the use of 2 Da for the precursor mass tolerance. Some search engines will model the potential for isotope cluster peaks explicitly, making it possible to specify a tight precursor mass tolerance and a small number of non-monoisotopic isotope cluster peaks to test, in addition to the monoisotopic mass of the peptide sequence. This is particularly appropriate for newer mass spectrometers, such as the Orbitrap, which can measure the precursor ion with high mass accuracy, and for which a precursor mass tolerance of 0.05 Da is appropriate. The isotope cluster precursor selection parameter is called #¹³C by Mascot and is generally set to a small number like 1 or 2.

It should be noted that peptide sequences may fail to match the mass of their experimental precursors for a variety of reasons. An incorrect charge state determination for an experimental precursor ion will make it impossible to match against its peptide sequence, while incorrect fixed or missing variable modifications will make it impossible for the *in silico* computation of a peptide mass to match the experimental value.

4.3 Execute Search

Once the spectra are prepared and uploaded and the search parameters are set, the execution of the search is generally straightforward.

4.4 Results Interpretation

We provide some general guidelines for results interpretation based on the following principles. First, peptide-spectrum match scores merely assess the strength of the fragment evidence and cannot establish the correctness of a match, so estimates of statistical significance must be used to control the number of false-positive peptide identifications, and second, due to the proteolytic digest of proteins to peptides in the shotgun/bottom-up proteomics workflow, proteins are not identified directly, but must instead be inferred from the peptide identification evidence.

4.4.1 Peptide-Spectrum Match Scores

Search engines compute a single overall score for each peptide-spectrum match to rank the peptides matched with each spectrum. The score is used to limit the number of retained peptide sequences per spectrum. Strong and weak scores for peptide-spectrum matches vary depending on the precursor ion's charge state, the quality of the spectrum, and the fragmentation characteristics of the (unknown) peptide represented by the precursor. We can assume that any good peptide-spectrum match score will usually give the best score to the correct peptide, if spectrum is of good quality. However, this does not imply that the best peptide identification for each spectrum is necessarily correct. If the score is sufficiently good, we may conclude the evidence for the peptide-spectrum match is strong, and the peptide is likely the correct identification. If the score is too weak, we must conclude that

the correct peptide cannot be determined from the peptide-spectrum match even if the (unknown) correct peptide has the best score. Various rule-of-thumb thresholds have been published for selecting the likely correct peptide-spectrum matches in the results from specific search engines, but these have proven less reliable than the statistical significance methods described below.

4.4.2 Peptide-Spectrum Match Characteristics

Peptide identifications may also have a variety of quantitative and qualitative values associated with the peptide-spectrum match. Common match characteristics include the experimental precursor m/z , experimental precursor mass, theoretical precursor mass, presumed charge state, missed cleavages, N- and C-terminal sequence proteolytic enzyme specificity, number of matching b - and y -ions, and the peptide rank. All of these match characteristics can be used to assess the quality of the match, though the search engines themselves do not usually factor these characteristics directly into their scores. These values can be used as additional filtering criteria, post-search, to preferentially remove false-positive identifications. Generally accepted characteristics of lower-quality identifications include semi-specific and nonspecific proteolytic enzyme N- or C-termini, number of missed cleavages, and the difference between the experimental and theoretical precursor mass. These match characteristics can also be used to evaluate instrument performance and sample preparation.

4.4.3 Peptide-Spectrum Match Statistical Significance

In the absence of a way to easily decide whether a specific rank 1 peptide identification should be accepted, various statistical significance techniques have been employed to provide search engines with calibrated scores. For a peptide-spectrum match of peptide P with spectrum S , the per spectrum p -value statistic assesses the *probability* that a *single random peptide* would score as well as, or better than, peptide P when matched against spectrum S . The E -value statistic assesses the *expected number of random peptides* that would score as well as, or better than, peptide P in a search of a sequence database of “random” peptides of the same size as the one actually searched. The E -value corrects for the increased number of false-positive identification at a given p -value when searching a large sequence database. The scores of random peptides matched with spectrum S are used to calibrate the range of poor scores for spectrum S , which hopefully does not include that of peptide P . Crudely, if the peptide’s score is not considerably better than those of random peptides, the evidence for its correctness is weak.

Unfortunately, each search engine uses a different notion of randomness in its p -value and E -value estimation techniques, which make E -values from different search engines difficult to compare. Well-estimated E -values, however, *do* provide a normalized score for comparisons of peptide-spectrum matches between spectra and peptides from the same analysis, even when the peptides are of

different lengths and the spectra vary in their signal and noise characteristics. E -values are always monotonic with the peptide-spectrum match scores for peptides matched to a specific spectrum, and so preserve rank. E -values are computed internally by the search engine and output, with the peptide-spectrum match score, as part of the search results.

As normalized, calibrated peptide-spectrum match scores, p -values and E -values make it possible to apply a single acceptance threshold to an entire set of rank 1 peptide identifications. However, determining a suitable E -value threshold for accepting peptide identifications is not straightforward, as filtering by the per spectrum E -value does not control the number or proportion of false peptide identifications in the accepted peptide identification set.

The false discovery rate (FDR) statistic measures the proportion of incorrect peptide identifications in any accepted peptide identification set, such as the set determined by thresholding on the scores, p -values, or E -values of rank 1 peptide identifications. To estimate the (spectral) FDR in practice, the spectra are also searched against a similarly sized sequence database of known false protein sequences (called decoys) and the same score, p -value, or E -value threshold applied to the decoy peptide identifications. Usually the decoy protein sequences are merely the reverse of the original (or target) protein sequences. The number of decoy peptide identifications that pass the threshold is taken as an estimate of the number of incorrect peptide identifications in the original accepted set of peptide identifications and FDR computed accordingly. The FDR statistic should only be computed using scores valid for comparing peptide identifications between spectra, so p -values and E -values should be used, if available. The estimated FDR can be computed for every possible E -value threshold to establish a relationship between E -value and FDR, making it possible to achieve a specific FDR target such as 1 %, 5 %, or 10 %.

Decoy-based FDR estimation can be readily carried out, regardless of a search engine's scoring scheme or its methodology for computing p -values and E -values. FDR estimates, computed consistently from the search results against a common target-decoy sequence database, can also be used to compare search engine sensitivity and specificity on a level playing field. The PepArML meta-search engine [20, 26] provides one such platform for comparing search engines with a consistently estimated FDR value. PepArML currently supports X!Tandem with native [18], k-score, and s-score scoring plug-ins [27] and OMSSA [28], MyriMatch [29], MSGF+ [19], and Mascot [16] search engines. Multiple search engines are most easily compared by using the q -value plot, which shows the relationship between the estimated FDR value and the number of spectra with peptide identifications in the corresponding accepted set. Any vertical line on the q -value plot indicates the number of

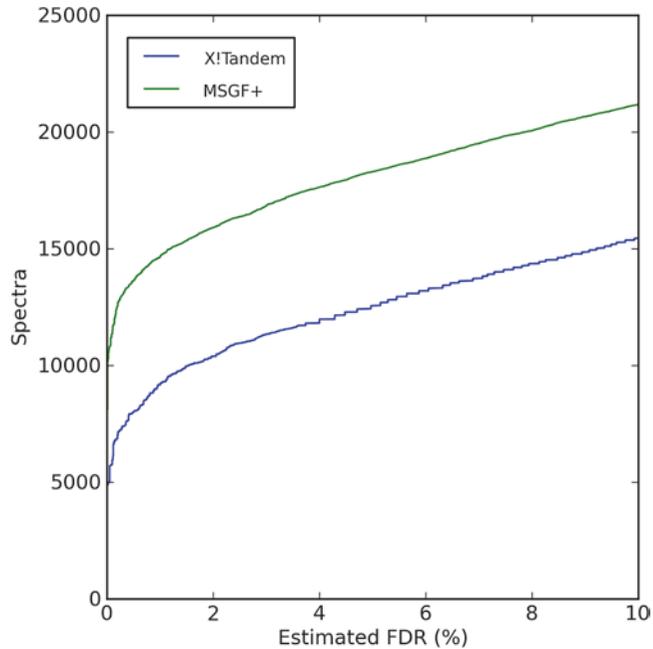


Fig. 5 q -value plot comparing X!Tandem and MSGF+ performance. Spectral data from Freese et al. [31]. Peptide identification searches carried out using PepArML [20]

spectra identified by a search engine at a given level of spectral FDR-based specificity. See Fig. 5 for an example of a q -value plot generated by PepArML.

4.4.4 Protein Identifications from Significant Peptides

With statistically significant peptide spectrum matches identified, it is now possible to characterize the protein content of the sample. Broadly, we look for proteins with significant peptide identifications to two or more distinct peptide sequences. While it is often tempting to accept proteins identified by peptide identifications to just a single peptide sequence with sufficiently small E -value, there are a variety of ways in which a false-positive peptide identification may be statistically significant, but nevertheless incorrect. Sequence homology is one common source of such errors. Requiring two or more distinct peptides per protein provides some measure of protection against these statistically significant, yet incorrect, peptide identifications.

For big spectral datasets with a large number of significant peptide identifications after filtering by a fixed FDR value, even a small *proportion* of incorrect identified peptides in the accepted set can result in enough false peptides that some incorrect proteins satisfy the two distinct peptide constraint. We can estimate the size of this effect using a simple binomial model based on the number of distinct decoy peptides n in the accepted set of peptide identifications

and the number of protein sequences N in the sequence database. Under the assumption that identified decoy peptides are distributed uniformly over the decoy sequences, the expected number of decoy proteins D with at least k peptides can be determined using the binomial distribution:

$$E[D] = N \Pr[X \geq k]; \quad X \sim \text{Binomial}\left(p = \frac{1}{N}, \text{ trials} = n\right) \quad (4)$$

Crucially, we point out that the expected number of decoy proteins passing the k distinct peptide constraint is dependent on the *absolute* number of identified decoy peptides, not the *proportion* of decoy peptides in the filtered peptide identification set. When the total number of accepted peptide identifications is large, the number of decoy peptides also increases, driving the expected number of decoy proteins up to a potentially unacceptable level. Figure 6 shows how the expected number of decoy proteins changes with the number of accepted decoy peptides according to the binomial model. Unfortunately, the number of decoy peptides at each E -value threshold is rarely reported and cannot be directly inferred from the spectra FDR rate. We point out that it is quite

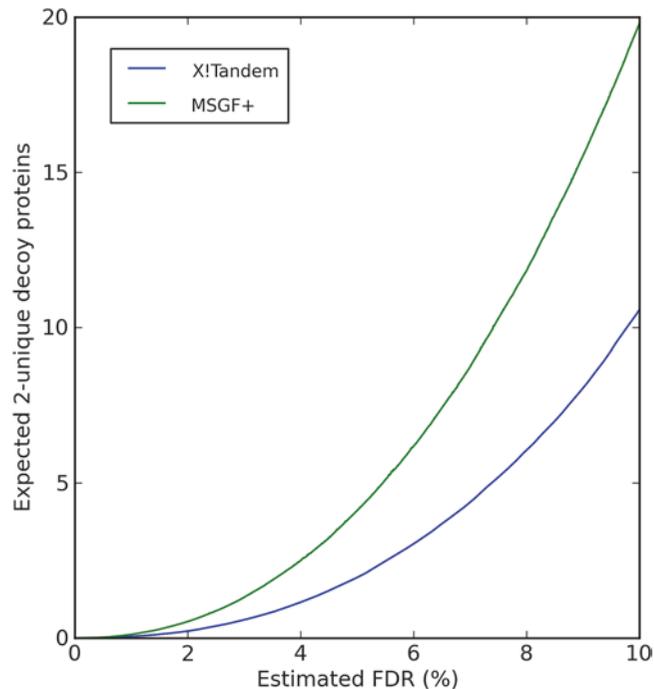


Fig. 6 Expected number of decoy proteins with at least two distinct peptides under the binomial model, for various spectral FDR cutoffs. Spectral data from Freese et al. [31]. Peptide identification searches carried out using PepArML [20]

feasible, when estimating the q -value curve, to establish the relationship between the number of decoy peptides and the spectral FDR and apply Eq. (4) to estimate the corresponding expected number of decoy proteins with at least k distinct identified peptides. By bounding the expected number of decoy proteins with at least two distinct peptides at a small value, say 0.05, we can establish a conservative spectral FDR filtering threshold, even for very large spectral datasets.

Multiple statistically significant peptide identifications to the same peptide should not increase confidence that the identifications are correct. These repeated identifications are often an artifact of instruments' precursor sampling strategies. MS/MS spectra of the same precursor ion are often acquired multiple times during its elution envelope, resulting in repeated, related spectra, which tend to have correlated errors. Correlated errors are also likely when multiple peptide ion charge states are seen.

Furthermore, while we usually assume that distinct peptides represent independent peptide identifications, there are occasions when this is not sufficient. Peptides with common N- or C-termini sequences are sometimes identified due to precursor charge state enumeration (*see* Subheading 4.1) or nonspecific proteolytic cleavage (*see* Subheading 4.2.4). These dependent identifications are an extremely common artifact of some spectral processing software, which enumerates identical MS/MS spectra with charges states 2+ and 3+. In the case of these dependent spectra too, only one peptide should be considered for the purposes of determining matched proteins.

A reasonable approach to conservative protein identification is to require at least two distinct peptide sequences per inferred protein after filtering the peptide-spectrum matches using a FDR-based threshold, chosen to ensure an appropriate protein false discovery rate. While the simple binomial model described above can be used to provide a general idea of appropriate thresholds to avoid a significant number of false-positive protein identifications, its assumptions are not a good match for the properties of real protein sequence databases, where proteins have varying lengths and share peptide sequences. Retaining decoy peptides throughout the protein inference procedure makes it possible to use the number of inferred decoy proteins as an (still rather imperfect) estimate of the likely number of incorrect proteins. Significantly, in large datasets which identify a significant proportion of the proteins in the sequence database, the protein FDR estimate should be adjusted to account for the possibility of false-positive peptide identifications falling on true positive proteins, using a technique such as that implemented in the MAYU [30] tool.

Table 3 shows protein inference results from the Freese et al. spectra dataset [31] for various spectral FDR filtering thresholds and two different search engines.

Table 3

Protein inference results from PSMLab prototype. Spectral data from Freese et al. [31]. Peptide identification searches carried out using PepArML [20]

Search	Spectral	Target	Decoy		MAYU
Engine	FDR (%)	Proteins	Proteins	FDR (%)	FDR (%)
X!Tandem	1	1256	0	0.00	0.00
X!Tandem	5	1477	18	1.22	1.20
X!Tandem	10	1662	69	4.15	4.08
MSGF+	1	1646	1	0.06	0.06
MSGF+	5	1855	27	1.46	1.43
MSGF+	10	2037	122	5.99	5.86

4.4.5 Protein Identification and Shared Peptides

Unfortunately, the strategies outlined in the previous section only control for the chance observation of false-positive proteins given FDR-based peptide identification filtering and at least two distinct peptides per protein. Another common source of incorrect protein identification in peptide identification analyses is due to peptide sequences found in both true-positive and false-positive proteins. Shared peptide sequences are common when searching a sequence database with protein sequences from many species due to sequence homology, but they are also observed in protein sequence databases containing sequence variants and protein isoforms. Thus, while statistically significant peptide identifications are *necessary* to conclude the presence of a protein, they are *not sufficient*. We can conclude that *at least one* of the proteins supported by statistically significant, but shared, peptide identification evidence is present—the difficulty is in determining which one.

When proteins with shared peptide sequences have multiple additional significant peptide identifications to peptide sequences that are not shared, the choice is clear and these should be retained. Proteins whose identified peptides are a strict subset of those of another protein should be eliminated. Proteins whose identified peptide are exactly equivalent and must be treated as equally valid conclusions—but at most one of these should be reported. Unfortunately, some tools choose between equivalent proteins arbitrarily, which can sometimes give the impression that an uncharacterized protein or unexpected protein isoform has been identified when, in fact, the identified peptide sequences can also be found in a well-understood, canonical protein form. Heuristics and keyword matching can sometimes be used to select the best representative from proteins with equivalent peptide identification evidence, but the informatician should always be on guard for

misleading representatives for proteins with the same set of identified peptides. The remaining shared peptide cases require the careful examination of distinguishing peptide identifications to determine if the evidence is strong enough to support the existence of multiple protein sequences. Usually this is not the case, and at most one sequence should be reported.

A good general rule of thumb is that each identified protein should have its own evidence, not shared with any other identified protein. The identification of proteins with shared peptides is fine, as long as each protein has peptides that are not shared. This permits both the canonical form of a protein and an isoform sequence to be identified, as long as the isoform sequence has sufficient peptide evidence not shared with the canonical form. If there is insufficient unshared evidence for the isoform sequence, then only the canonical sequence should be reported.

With the issue of shared peptides in mind, then, we strengthen the above guidance to require at least two distinct (and nonoverlapping) peptide sequences per inferred protein, not shared with any other inferred protein, after filtering the peptide-spectrum matches using a FDR-based threshold, chosen to ensure an appropriate protein false discovery rate.

4.4.6 Protein Identification and Protein Parsimony

Protein inference algorithms which find the smallest set of protein sequences sufficient to explain a given accepted set of identified peptides implement a principle known as parsimony. Protein parsimony algorithms must necessarily remove any protein whose peptides are equivalent to, or a subset of, the peptides of another protein, thereby removing redundant protein identifications due to shared peptides. On the other hand, the traditional protein parsimony approach will retain a protein sequence even if it has just a single peptide sequence of its own. This violates the principles outlined in Subheading 4.4.4. Unfortunately, very few protein inference tools satisfy both Subheadings 4.4.4 and 4.4.5, and the informatician is left to apply manual or heuristic filtering to the protein inference results as needed. The Edwards lab is developing a peptide identification post-processing toolkit, called PSMLab, which satisfies both of these principles. The results shown in Table 3 represent the output of a PSMLab prototype.

References

1. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422(6928):198–207
2. McDonald WH, Yates JR (2003) Shotgun proteomics: integrating technologies to answer biological questions. *Curr Opin Mol Ther* 5(3):302–309
3. Sadygov RG, Cociorva D, Yates JR (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 1(3):195–202
4. Johnson R, Davis M, Taylor J et al (2005) Informatics for protein identification by mass spectrometry. *Methods* 35(3):223–236

5. Maccoss M (2005) Computational analysis of shotgun proteomics data. *Curr Opin Chem Biol* 9(1):88–94
6. Nesvizhskii AI (2007) Mass spectrometry data analysis in proteomics, *Methods in Molecular Biology*, vol 367, Humana Press, chap Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching, pp 87–119
7. Deutsch EW, Lam H, Aebersold R (2008) Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol Genomics* 33(1):18–25
8. Taylor A, Johnson RS (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 11:1067–1075
9. Chen T, Kao MY, Tepel M et al (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 8(3):325–337
10. Bafna V, Edwards N (2003) On de novo interpretation of tandem mass spectra for peptide identification. In: RECOMB '03: Proceedings of the seventh annual international conference on research in computational molecular biology. ACM Press, pp 9–18
11. Frank A, Pevzner P (2005) Pepnovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal Chem* 77(4):964–973
12. Mann M, Wilm M (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 66(24):4390–4399
13. Tanner S, Shu H, Frank A et al (2005) Inspect: identification of post translationally modified peptides from tandem mass spectra. *Anal Chem* 77(14):4626–4639
14. Tabb DL, Ma ZQ, Martin DB et al (2008) DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* 7(9):3838–3846
15. Dass C (2001) Principles and Practice of Biological Mass Spectrometry. John Wiley & Sons Inc.
16. Perkins DN, Pappin DJ, Creasy DM et al (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–3567
17. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5:976–989
18. Craig R, Beavis RC (2004) Tandem: matching proteins with tandem mass spectra. *Bioinformatics* 20:1466–1467
19. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 5:5277
20. Edwards NJ (2013) PepArML: a meta-search peptide identification platform for tandem mass spectra. *Curr Protoc Bioinformatics* 44(13):23.1–2323
21. Slagel J, Mendoza L, Shteynberg D et al (2015) Processing shotgun proteomics data on the amazon cloud with the Trans-Proteomic pipeline. *Mol Cell Proteomics* 14(2):399–404
22. The UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res* 38(Database Issue):D142–D148
23. Edwards NJ (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol Syst Biol* 3(102)
24. Keller A, Eng J, Zhang N et al (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 1(17)
25. Kessner D, Chambers M, Burke R et al (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24(21):2534–2536
26. Edwards N, Wu X, Tseng CW (2009) An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. *Clin Proc* 5(1)
27. MacLean B, Eng JK, Beavis RC et al (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* 22(22):2830–2832
28. Geer LY, Markey SP, Kowalak JA et al (2004) Open mass spectrometry search algorithm. *J Proteome Res* 3:958–964
29. Tabb DL, Fernando CG, Chambers MC (2007) Myrimatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6(2):654–661
30. Reiter L, Claassen M, Schrimpf SP et al (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* 8(11):2405–2417
31. Frese CK, Altelaar AFM, Hennrich ML et al (2011) Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-orbitrap velos. *J Proteome Res* 10(5):2377–2388

Bioinformatics Analysis of Top-Down Mass Spectrometry Data with ProSight Lite

Caroline J. DeHart, Ryan T. Fellers, Luca Fornelli, Neil L. Kelleher, and Paul M. Thomas

Abstract

Traditional bottom-up mass spectrometry-based proteomics relies on the use of an enzyme, often trypsin, to generate small peptides (typically < 25 amino acids long). In top-down proteomics, proteins remain intact and are directly measured within the mass spectrometer. This technique, while inherently simpler than bottom-up proteomics, generates data which must be processed and analyzed using software tools “purpose-built” for the job. In this chapter, we will show the analysis of intact protein spectra through deconvolution, deisotoping, and searching with ProSight Lite, a free, vendor-agnostic tool for the analysis of top-down mass spectrometry data. We will illustrate with two examples of intact protein fragmentation spectra and discuss the iterative use of the software to characterize proteoforms and discover the sites of post-translational modifications.

Key words Top-down, Mass spectrometry, Proteomics, ProSight Lite, Intact protein, Bioinformatics

1 Introduction

Complementing the speed and sensitivity of bottom-up proteomics [1, 2], top-down proteomics [3] offers a comprehensive view of proteoforms present in the sample (for a more detailed description of a proteoform, *see Note 1*) [4]. It is important to distinguish between top-down mass spectrometry and top-down proteomics. In top-down mass spectrometry, one or a few proteoforms are isolated and studied. This technique is typified by studies of histones [5–7] and immunoglobulins [8–10], among others, including the routine analysis of small (< 40 kDa) proteins. In top-down proteomics, a whole, unknown proteome is typically isolated, fractionated, and analyzed by mass spectrometry [11–13]. The tools used for the study of top-down mass spectrometry and top-down proteomics are related, but fundamentally different. Top-down proteomics requires searching tandem MS data against a database of known proteins and proteoforms [14, 15], while top-down mass

spectrometry is simpler. In the latter case, the experimenter generally knows the protein under study, but may want to characterize the location of post-translational modifications [5] or to better understand a new protein fragmentation technique [16]. Here we illustrate, with real data use cases, how to deconvolute and deisotope top-down mass spectrometry data with QualBrowser and Xtract (ThermoFisher, San Jose, CA) followed by analysis with ProSight Lite [17], a free tool for intact protein data analysis.

2 Materials

1. ProSight Lite [17] downloaded from <http://prosightlite.northwestern.edu>. This must be installed on a Microsoft Windows computer with .NET Framework 4.5.1 or greater.
2. Top-down mass spectrometry data. A ThermoFisher .raw file for a top-down standard used in this study has been provided at <http://hdl.handle.net/2022/20816>. In addition to the raw data, the output from Subheading 3.1 (below) is provided as PostXtract .raw files and text files containing the deconvoluted and deisotoped masses for use as input in Subheadings 3.2 and 3.3. A brief description of how the sample was prepared is found in **Note 2**.
3. Internet access to <http://www.uniprot.org>. This is required to automatically extract protein sequence information from a UniProt [18] entry.
4. (Optional). Software tools to view, deconvolute, and deisotope mass spectra. Many vendors supply a proprietary algorithm with their instruments. In this method, Xtract was used within QualBrowser from ThermoFisher Scientific. Other free options for the deconvolution and deisotoping of top-down mass spectrometry data include MS-Deconv [19] (available at <http://bix.ucsd.edu/projects/msdeconv/>) or YADA [20] (available at <http://fields.scripps.edu/yada/>); however, a full discussion of their use is beyond the scope of this chapter.
5. (Optional). Software tools to convert mass spectrometry data into an extensible markup language format such as .mzML and .mzXML. This may be required by deconvolution/deisotoping tools. One suggestion would be ProteoWizard [21], found at <http://proteowizard.sourceforge.net/downloads.shtml>.

3 Methods

3.1 Deconvolution and Deisotoping of Raw Data

1. For this exercise, download **TopDownStandard.raw** from the repository and open the file in QualBrowser. We will deconvolute and deisotope data for two of the five proteins present in the file, ubiquitin and carbonic anhydrase, as shown in Fig. 1

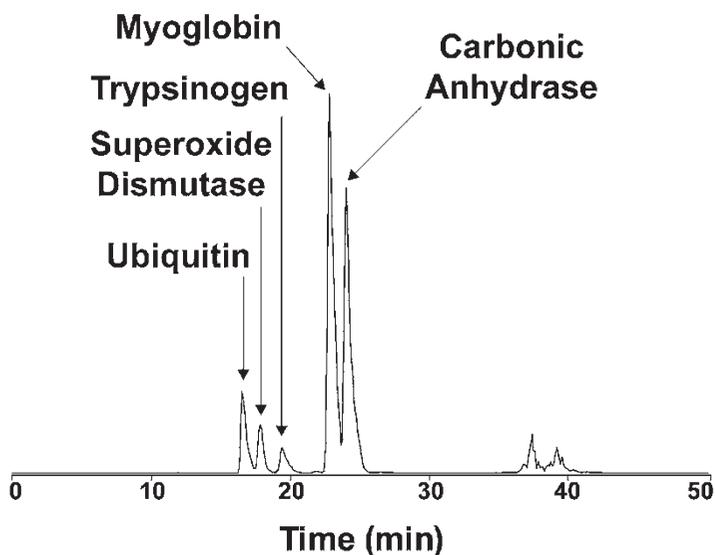


Fig. 1 Example chromatogram showing the separation of five intact proteins present in TopDownStandard.raw. The MS and MS/MS spectra for ubiquitin and carbonic anhydrase were chosen near the apex of the respective peaks

(see **Note 2** for more information). The goal of the steps to follow is to generate zero-charge, monoisotopic mass spectra for the precursor and fragmentation scans. Instructions presented below will work with ThermoFisher .raw files. Other workflows can be generated with other vendor-provided software or using openly available tools as described in **Materials**. All output files from the steps below are published with the dataset as indicated above in Subheading 2, **item 2**.

2. To prepare data for ProSight Lite, we require the intact mass (MS1 data) for each protein and a list of MS2 fragments observed. In QualBrowser, display the intact mass spectrum for ubiquitin from scan 334 (for more in-depth explanations of QualBrowser commands, see <http://planetorbitrap.com/download.php?filename=524212aa70353.pdf>).
3. Display the Xtract dialog for scan 334 by pinning the mass spectrum tab. Right click to bring up the options menu, select the **Export** menu, and then click **Xtract**. If **Xtract** is not available, see **Note 3**.
4. Set the Xtract parameters as displayed in Fig. 2. For top-down mass spectrometry, monoisotopic, zero-charge data are typically required. Further explanation of these parameters may be found in **Note 4**. Select Xtract to generate a deisotoped, deconvoluted spectrum from scan 334.
5. Select the new window that appears by pinning it. Navigate to scan number two of the PostXtract scan. This scan contains the monoisotopic, zero-charge ($^{13}\text{C}_0$) mass spectrum. Scan 1

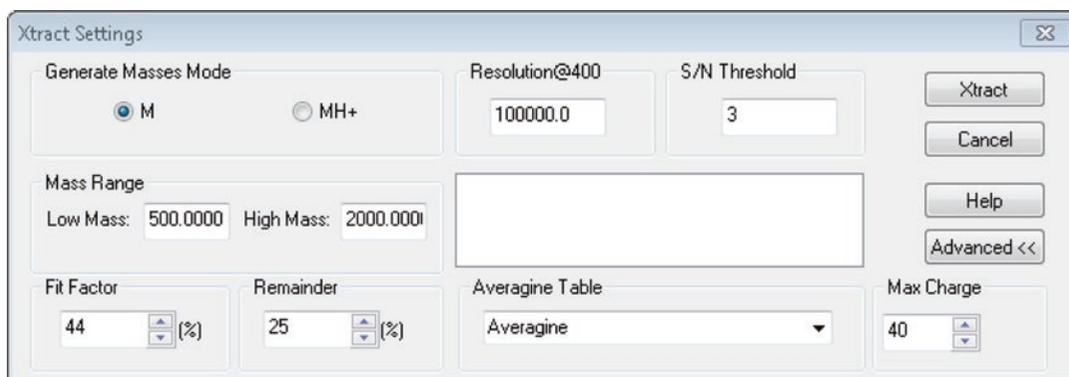


Fig. 2 Selected Xtract parameters to be used for deconvolution and deisotoping. An explanation of these parameters may be found in **Note 4**

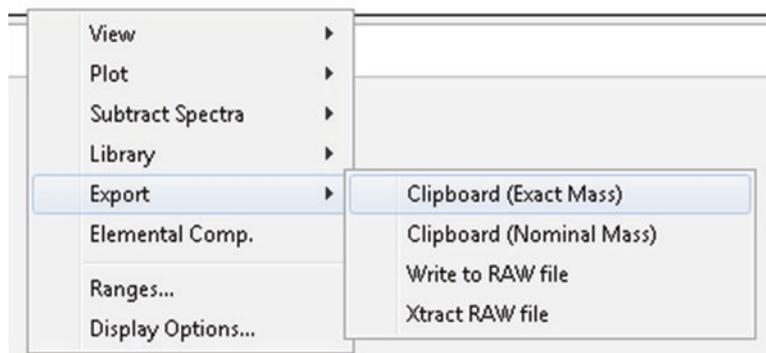


Fig. 3 Demonstration of exporting post-Xtract data onto the clipboard for further analysis by ProSight Lite

contains the zero-charge mass spectrum that has not been deisotoped (and thus contains all of the isotopes). Scan 3 contains the m/z values from the input spectrum that were deconvoluted and deisotoped by the Xtract algorithm. Scan 4 contains the remaining m/z values from the input spectrum that were *not* deconvoluted and deisotoped by the Xtract algorithm.

6. Copy the monoisotopic, zero-charge ($^{13}\text{C}_0$) mass spectrum to the clipboard. Right click to bring up the options menu, select the **Export** menu, and then click **Clipboard (Exact Mass)** as shown in Fig. 3.
7. Open a spreadsheet program such as Microsoft Excel and paste the data from the clipboard. Note the most intense intact mass: **8,559.6556 Da**. Save the file as **UbiqMS1.xlsx**.
8. Repeat steps 3–7 with scan 330. Scan 330 is an HCD fragmentation scan of the 11+ charge state of ubiquitin. Save the resulting file as **UbiqMS2.xlsx**.

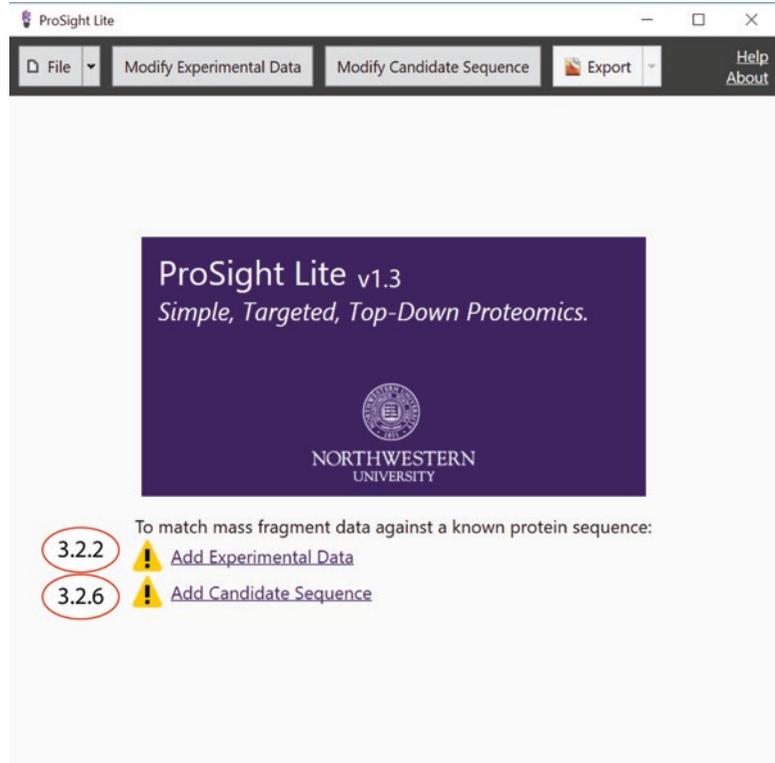


Fig. 4 Home screen of ProSight Lite showing the two pieces of information needed to perform the analyses outlined in Subheading 3.2, **steps 2 and 6**

9. Repeat steps 3–7 with scan 556. Scan 556 is an intact mass spectrum of bovine carbonic anhydrase. Note the most intense intact mass: **29,006.8254 Da**. Save the resulting file as **CAMS1.xlsx**.
10. Repeat steps 3–7 with scan 561. Scan 561 is an HCD fragmentation scan of the 32+ charge state of carbonic anhydrase. Save the resulting file as **CAMS2.xlsx**.

3.2 ProSight Lite Analysis of Bovine Ubiquitin (Simple Analysis Case)

1. Open ProSight Lite and note the two pieces of information required: experimental data (MS1 and MS2) and sequence information (Fig. 4).
2. Click **Add Experimental Data**. Note the new window that appears (Fig. 5).
3. In **Precursor**, enter **8,559.6556** as determined in Subheading 3.1, step 7, above or copy from **UbiqMS1.xlsx**.
4. Open **UbiqMS2.xlsx** and copy the mass data from Column A. Paste these values into the **Fragments (one per line)** text box.

Experimental Data

Precursor **3.2.3**
8559.6556

Precursor Mass Type
 Monoisotopic Average

Fragments (one per line) **3.2.4**

6110.524220
6127.352323
6210.369302
6228.374662
6325.40033
6343.424054
6395.409815
6412.440663
6422.442228
6430.439334
6440.426843
6468.450974
6491.460719
6509.491103
6527.50624
6620.52017
6638.546767

Mass Mode **3.2.5**
 M (neutral) MH+

Fragmentation Methods **Note 5**

CID
 HCD
 SID
 ECD
 ETD
 EThcD
 IRMPD
 UVPD
 BY and CZ*

Fragmentation Tolerance
10 ppm

Save Cancel

Fig. 5 Experimental Data panel. Precursor and fragmentation data must be provided and other parameters must be set prior to further analysis by ProSight Lite

5. Select the following parameters (*see* **Note 5** for more detailed descriptions), then click **Save**:
 - (a) Precursor Mass Type: **Monoisotopic**
 - (b) Mass Mode: **M (neutral)**
 - (c) Fragmentation Method: **HCD**
 - (d) Fragmentation Tolerance: **10 ppm**
6. Click on **Add Candidate Sequence**. Note that a new window appears (Fig. 6).
7. In the sequence text box, enter the sequence of bovine ubiquitin: MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPDQQLRIFAGKQLEDGRTLSDYNIQKESTLHLVLRLLRGG
8. Select **No Modification** for cysteine and methionine, and then click **Save**.
9. Results are obtained (Fig. 7) indicating a strong agreement between the experimental data and the candidate sequence. The P-score is 2.5×10^{-68} , multiple fragments cover both termini, and the intact mass differs from the theoretical by 0.04 Da or < 5 ppm. Flags pointing to the upper left indicate

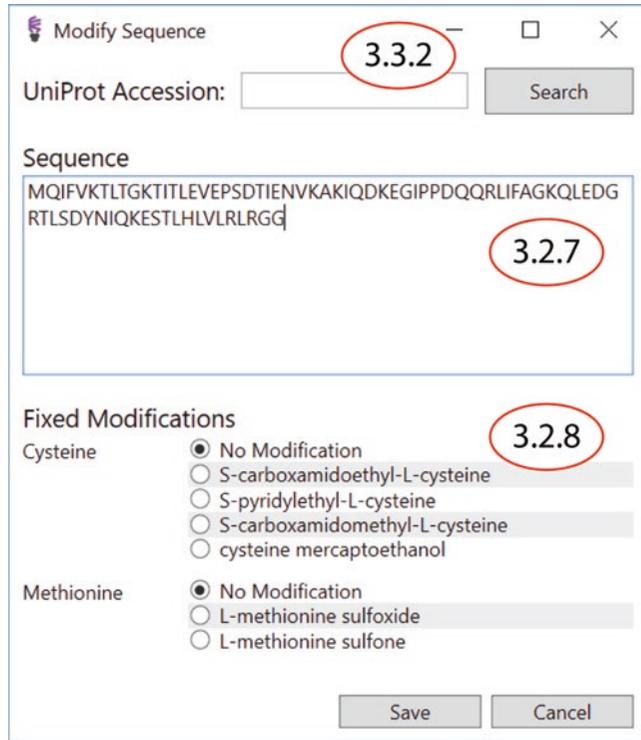


Fig. 6 Modify Sequence panel. An amino acid sequence may be directly pasted into the sequence text box or automatically retrieved using a UniProt accession number

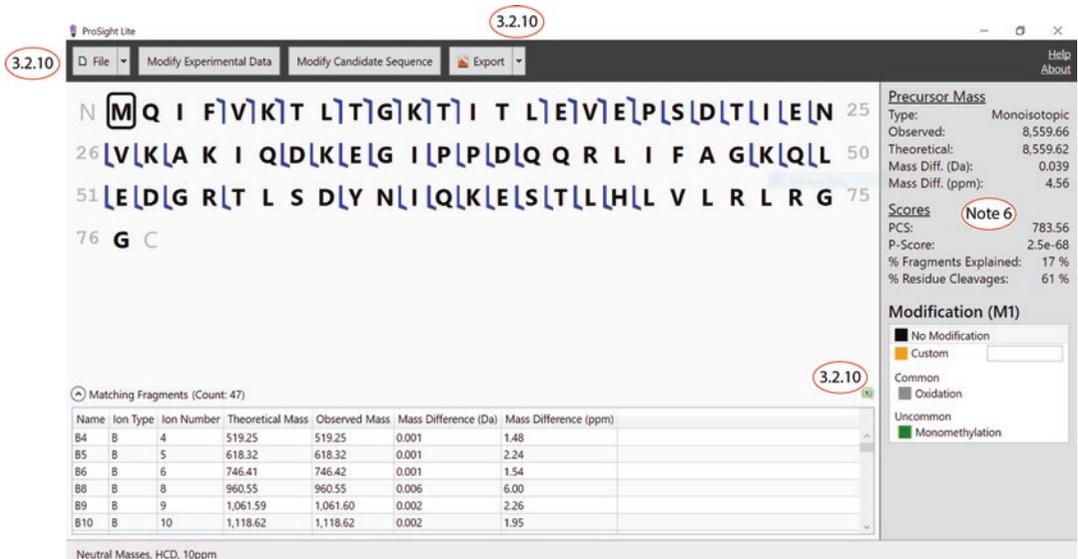


Fig. 7 Successful analysis of bovine ubiquitin with ProSight Lite. Scores are discussed in detail in **Note 6**

N-terminal fragments, while flags pointing to the lower right indicate C-terminal fragments. The fragment color is determined by the Roepstorff fragment type [22]: a/x are green, b/y are blue, and c/z are red. Scores are further described in **Note 6**.

10. Exports:

- (a) A list of matching fragment ions and their mass errors can be downloaded by selecting the Microsoft Excel icon to the upper right of the matching fragment table.
- (b) Fragmentation maps may be downloaded in PNG (raster graphic) or SVG (vector graphic) under the **Export** menu.
- (c) Results may be saved for future use or shared with collaborators using the PCML format under the **File... Save** menu.

**3.3 ProSight Lite
Analysis of Bovine
Carbonic Anhydrase II
(Iterative Analysis
Case)**

1. Open a new instance of ProSight Lite and **Add Experimental Data** as in Subheading 3.2, step 2–5, above with the following changes:
 - (a) Use data from **CAMS1.xlsx** in Subheading 3.2, step 3 (either enter **29,006.8254** as determined in Subheading 3.1, step 9, above or copy from the file).
 - (b) Use data from **CAMS2.xlsx** in Subheading 3.2, step 4.
2. Click on **Add Candidate Sequence**. In the UniProt Accession text box (Fig. 6), enter **P00921**, and click **Search**. This populates the sequence text box with the amino acid sequence for bovine carbonic anhydrase II. Select **No Modification** for cysteine and methionine, and then click **Save**. This step prevents errors in the process of copying and pasting data from multiple sources and should be used when the protein being studied is present in the UniProt Knowledgebase.
3. Iteration 1. Results are obtained (Fig. 8) indicating poor agreement between the experimental data and the candidate sequence. The excellent P-score of 2.7×10^{-20} arising from the strong coverage of the C-terminal region of the protein indicates that the protein is identified as carbonic anhydrase II; however, the theoretical intact mass is 88.89 Da higher than the observed intact mass, and no N-terminal fragment ions are observed. At this stage, we have identified the protein, but have not fully characterized the proteoform due to the large mass error.
4. Iteration 2. To correct this mismatch between the theoretical and observed data, first select **Modify Candidate Sequence**, remove the N-terminal methionine from the sequence in the text box, and click **Save**. The results (Fig. 9) show an identical P-score (2.7×10^{-20}) and no new N-terminal fragment ions; however, the mass difference shows that the theoretical mass is now 42.15 Da lower than the observed mass.

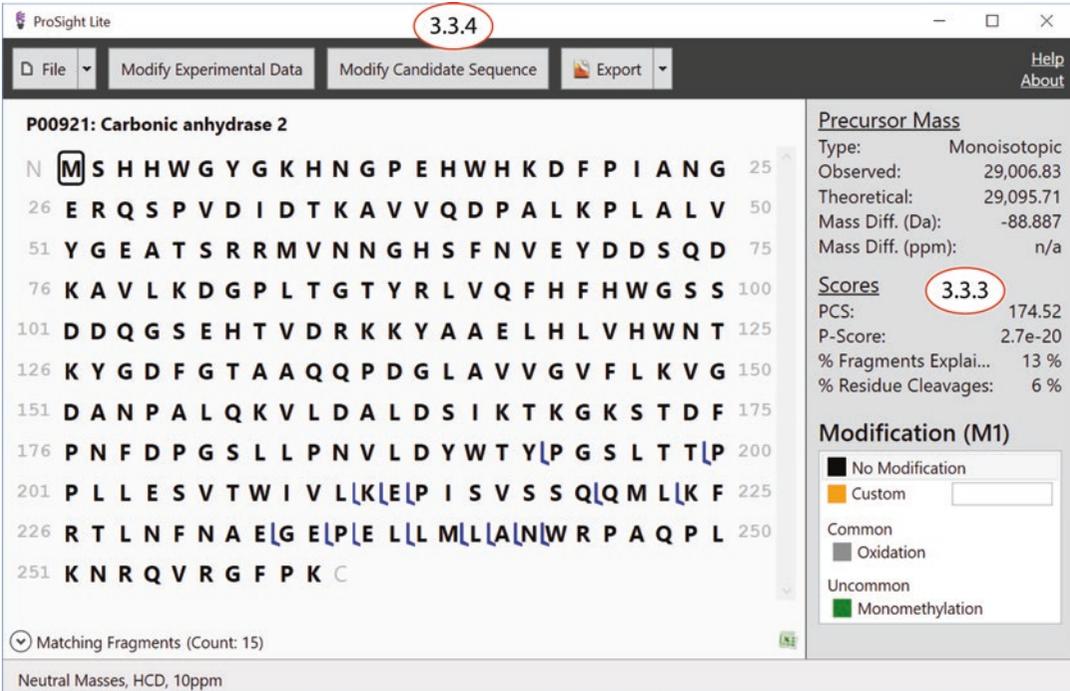


Fig. 8 Iteration 1: Partial characterization of bovine carbonic anhydrase. The protein is identified, but the proteoform (*see Note 1* for a detailed description) is not fully characterized

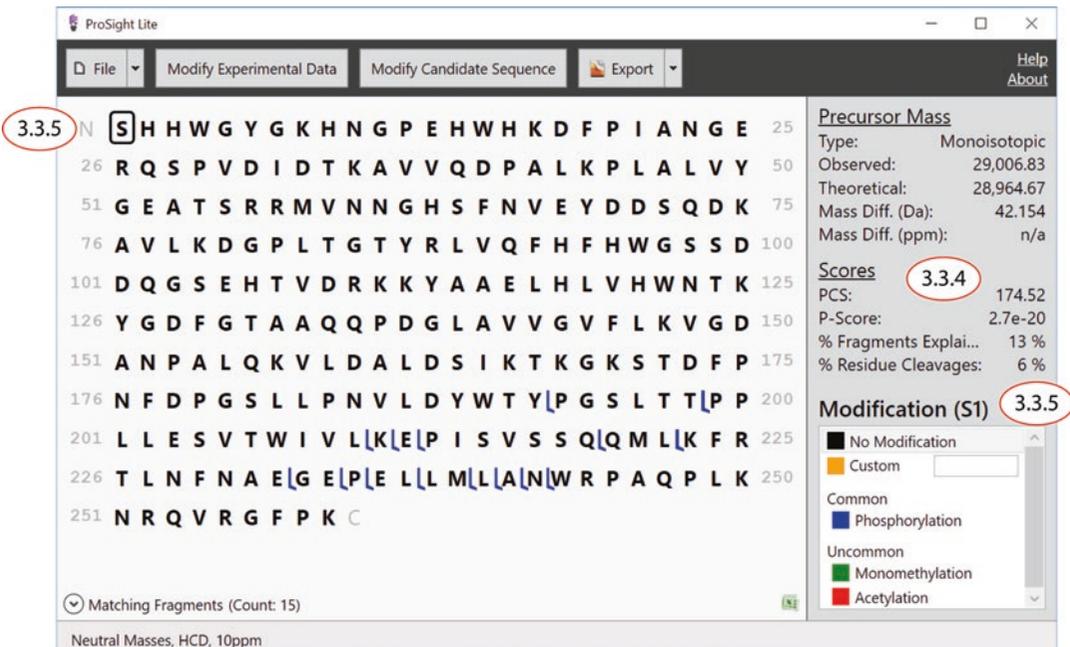


Fig. 9 Iteration 2: Improved characterization of bovine carbonic anhydrase as a result of removing the initial methionine from the sequence prior to re-searching

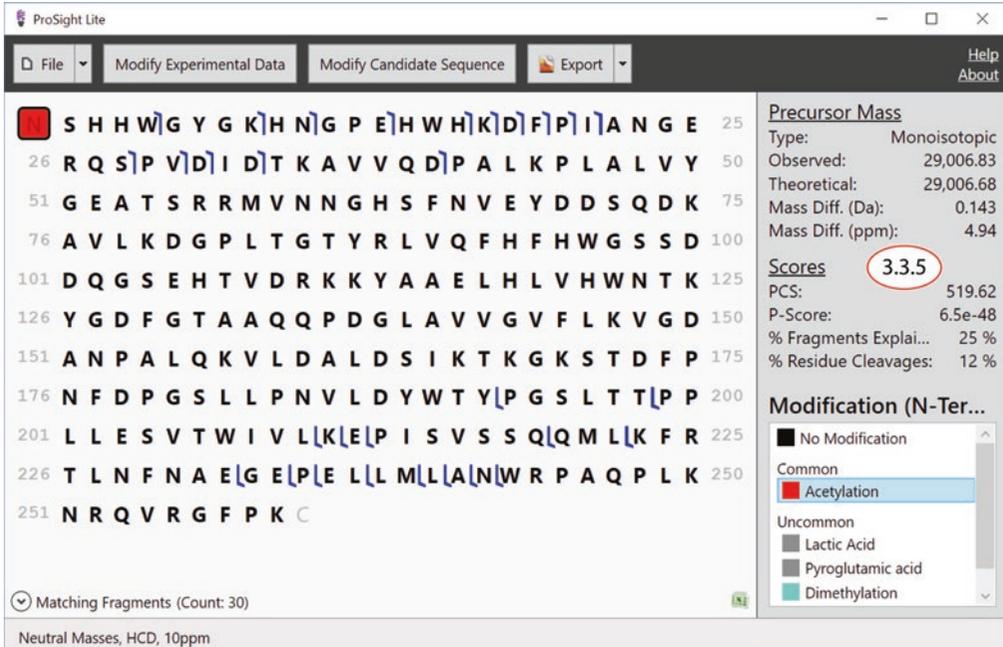


Fig. 10 Iteration 3: Complete characterization of the N-terminal methionine cleaved, N-terminally acetylated bovine carbonic anhydrase II proteoform

5. Iteration 3. The final iteration requires the modification of the N-terminus with an acetyl group (with a mass of 42.01 Da). To add this modification, click on the gray “N” to the left of the sequence and, from the **Modification** box on the lower right, select “**Acetylation**.” Note that once changes are made, the coverage and scores update in real time. This feature allows for many changes to be made iteratively as the user improves the match of the theoretical proteoform to the observed data. The final result can be seen in Fig. 10. The P-score is improved to 6.5×10^{-48} , multiple fragment ions report on both termini, and the precursor mass error is reduced to just 0.14 Da or 4.9 ppm. After iteration 3, the proteoform has been correctly and fully characterized.
6. Exports can again be performed as described above in Subheading 3.2, step 10.

4 Notes

1. A proteoform is a single instance of a group of protein products arising from a specific gene incorporating multiple sources of potential variation including alternative splicing, single-nucleotide polymorphism, and post-translational modifications, among others [4].

2. The top-down standard is a solution containing bovine carbonic anhydrase (Sigma C2624), equine myoglobin (Sigma M5696), bovine trypsinogen (Sigma T1143), and bovine ubiquitin (Sigma U6253). A nanoLCMS run is performed in which a sample containing 0.6 pmol carbonic anhydrase, 1.1 pmol myoglobin, 0.5 pmol trypsinogen, and 0.1 pmol ubiquitin is injected onto a 75 μm ID \times 15 cm PLRP-S column (5 μm d_p , Agilent) and eluted from 5 % acetonitrile in water + 0.1 % formic acid to 52 % acetonitrile in water + 0.1 % formic acid over 27 min at a flowrate of 300 nL/min. Data are acquired in data-dependent MS/MS mode on a ThermoFisher Velos Orbitrap Elite targeting the top 1 m/z value. That species is isolated in a 15 m/z window and HCD fragmentation occurs at 25 % NCE. Bovine superoxide dismutase (SOD1) is present as a contaminant in bovine carbonic anhydrase.
3. If Xtract is not available within your copy of QualBrowser, contact your local ThermoFisher mass spectrometry sales representative to find out how to enable this feature.
4. Description of Xtract parameters with QualBrowser:
 - (a) **Generate Masses Mode:** Specifies whether Xtract returns zero-charge (**M**) or protonated, singly charged (**M+**) data. In this chapter, we use **M** as the convention as it is more intuitive.
 - (b) **Mass Range:** Specifies the m/z range to be deconvoluted and deisotoped. This is typically set to the start m/z and end m/z of the scan being processed.
 - (c) **Resolution@400:** Specifies the mass resolving power ($m/\Delta m$) of the measured scan. This parameter is only used as a default if the resolving power cannot be determined from the data.
 - (d) **S/N Threshold:** Specifies the lowest signal-to-noise ratio that the Xtract algorithm will confer with trying to deconvolute and deisotope the mass spectrum. Altering the threshold lower will generate more masses; however, many of these masses arise from noise and may be incorrect. A higher S/N threshold will generate fewer noise-related masses, but also fewer real masses.
 - (e) **Fit Factor:** Specifies the minimum fit parameter used by the Xtract algorithm. Increasing this parameter requires a more perfect fit of the data to the theoretical spectrum to be considered real; reducing this parameter requires a poorer fit to be considered real.
 - (f) **Remainder:** Specifies the remainder of the fit that is left in the scan during Xtract analysis. This remainder is used to deconvolute overlapping, lower-intensity ion signals.

- (g) **Average Table**: Specifies the model for the isotopic abundances to be used by Xtract. This should not be changed during typical operation.
 - (h) **Max Charge**: Specifies the maximum charge to be considered by the Xtract algorithm. This should be set at or above the highest charge expected in the mass spectrum.
5. Description of the experimental data parameters for ProSight Lite.
- (a) **Precursor Mass Type** is derived from the Xtract results. Selecting scan 2 in Subheading 3.1, step 5, generates monoisotopic masses.
 - (b) **Mass Mode** is derived from the Xtract parameters; *see Note 4*.
 - (c) **Fragmentation Method** is determined by the fragmentation used in the mass spectrometer instrument method.
 - (d) **Fragmentation Tolerance** is intrinsic to your instrument's mass measurement accuracy. Setting this parameter too narrow will discard many real masses; however, setting this parameter too wide will negatively affect the confidence in the results [23].
6. Description of scores in ProSight Lite:
- (a) **PCS**. Proteoform characterization score described in previous manuscript [6]. Range, 0–infinity. Higher numbers indicate better scores.
 - (b) **P-Score**. ProSight P-Score described in previous manuscript [23]. Range, 0–1. Lower numbers indicate better scores.
 - (c) **% Fragments Explained**. The percentage of the total number of fragments observed that are explained by the fragmentation data. This percentage can vary depending on the quality of deconvolution/deisotoping algorithm. Range: 0–100 %. Higher numbers indicate better deconvolution (less noise); lower numbers either indicate that more than one protein was isolated and fragmented simultaneously or that the deconvolution algorithm generated unrelated noise.
 - (d) **% Residue Cleavages**. The percentage of inter-residue bonds where at least one fragment ion was observed. Range: 0–100 %. Higher numbers indicate better scores.

Acknowledgments

The authors are grateful to members of the Kelleher Research Group/Proteomics Center of Excellence for helpful discussions, particularly Joseph Greer, Richard LeDuc, and Bryan Early.

This work was supported by Award No. P41GM108569 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Erickson BK, Jedrychowski MP, McAlister GC, Everley RA, Kunz R, Gygi SP (2015) Evaluating multiplexed quantitative phosphopeptide analysis on a hybrid quadrupole mass filter/linear ion trap/orbitrap mass spectrometer. *Anal Chem* 87(2):1241–1249. doi:[10.1021/ac503934f](https://doi.org/10.1021/ac503934f)
2. Beck S, Michalski A, Raether O, Lubeck M, Kaspar S, Goedecke N, Baessmann C, Hornburg D, Meier F, Paron I, Kulak NA, Cox J, Mann M (2015) The impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol Cell Proteomics* 14(7):2014–2029. doi:[10.1074/mcp.M114.047407](https://doi.org/10.1074/mcp.M114.047407)
3. Kelleher NL (2004) Top-down proteomics. *Anal Chem* 76(11):197A–203A
4. LM S, NL K, Consortium for Top Down P (2013) Proteoform: a single term describing protein complexity. *Nat Methods* 10(3):186–187. doi:[10.1038/nmeth.2369](https://doi.org/10.1038/nmeth.2369)
5. Zheng Y, Fornelli L, Compton PD, Sharma S, Canterbury J, Mullen C, Zabrouskov V, Fellers RT, Thomas PM, Licht JD, Senko MW, Kelleher NL (2015) Unabridged analysis of human histone H3 by differential top-down mass spectrometry reveals hypermethylated proteoforms from MMSET/NSD2 overexpression. *Mol Cell Proteomics*. doi:[10.1074/mcp.M115.053819](https://doi.org/10.1074/mcp.M115.053819)
6. Dang X, Scotcher J, Wu S, Chu RK, Tolic N, Ntai I, Thomas PM, Fellers RT, Early BP, Zheng Y, Durbin KR, Leduc RD, Wolff JJ, Thompson CJ, Pan J, Han J, Shaw JB, Salisbury JP, Easterling M, Borchers CH, Brodbelt JS, Agar JN, Pasa-Tolic L, Kelleher NL, Young NL (2014) The first pilot project of the consortium for top-down proteomics: a status report. *Proteomics* 14(10):1130–1140. doi:[10.1002/pmic.201300438](https://doi.org/10.1002/pmic.201300438)
7. Tian Z, Tolic N, Zhao R, Moore RJ, Hengel SM, Robinson EW, Stenoien DL, Wu S, Smith RD, Pasa-Tolic L (2012) Enhanced top-down characterization of histone post-translational modifications. *Genome Biol* 13(10):R86. doi:[10.1186/gb-2012-13-10-r86](https://doi.org/10.1186/gb-2012-13-10-r86)
8. Barnidge DR, Dasari S, Botz CM, Murray DH, Snyder MR, Katzmann JA, Dispenzieri A, Murray DL (2014) Using mass spectrometry to monitor monoclonal immunoglobulins in patients with a monoclonal gammopathy. *J Proteome Res* 13(3):1419–1427. doi:[10.1021/pr400985k](https://doi.org/10.1021/pr400985k)
9. Fornelli L, Damoc E, Thomas PM, Kelleher NL, Aizikov K, Denisov E, Makarov A, Tsybin YO (2012) Analysis of intact monoclonal antibody IgG1 by electron transfer dissociation Orbitrap FTMS. *Mol Cell Proteomics* 11(12):1758–1767. doi:[10.1074/mcp.M112.019620](https://doi.org/10.1074/mcp.M112.019620)
10. Zhang H, Cui W, Gross ML (2014) Mass spectrometry for the biophysical characterization of therapeutic monoclonal antibodies. *FEBS Lett* 588(2):308–317. doi:[10.1016/j.febslet.2013.11.027](https://doi.org/10.1016/j.febslet.2013.11.027)
11. Ansong C, Wu S, Meng D, Liu X, Brewer HM, Deatherage Kaiser BL, Nakayasu ES, Cort JR, Pevzner P, Smith RD, Heffron F, Adkins JN, Pasa-Tolic L (2013) Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. *Proc Natl Acad Sci U S A* 110(25):10153–10158. doi:[10.1073/pnas.1221210110](https://doi.org/10.1073/pnas.1221210110)
12. Cheon DH, Nam EJ, Park KH, Woo SJ, Lee HJ, Kim HC, Yang EG, Lee C, Lee JE (2015) Comprehensive analysis of low-molecular-weight human plasma proteome using top-down mass spectrometry. *J Proteome Res*. doi:[10.1021/acs.jproteome.5b00773](https://doi.org/10.1021/acs.jproteome.5b00773)
13. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480(7376):254–258. doi:[10.1038/nature10575](https://doi.org/10.1038/nature10575)
14. Liu X, Hengel S, Wu S, Tolic N, Pasa-Tolic L, Pevzner PA (2013) Identification of ultramodified proteins using top-down tandem mass spectra. *J Proteome Res* 12(12):5830–5838. doi:[10.1021/pr400849y](https://doi.org/10.1021/pr400849y)
15. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL (2007) ProSight PTM 2.0: improved protein identification and

- characterization for top down mass spectrometry. *Nucleic Acids Res* 35(Web Server issue):W701–W706. doi:[10.1093/nar/gkm371](https://doi.org/10.1093/nar/gkm371)
16. Cannon JR, Holden DD, Brodbelt JS (2014) Hybridizing ultraviolet photodissociation with electron transfer dissociation for intact protein characterization. *Anal Chem* 86(21):10970–10977. doi:[10.1021/ac5036082](https://doi.org/10.1021/ac5036082)
 17. Fellers RT, Greer JB, Early BP, Yu X, LeDuc RD, Kelleher NL, Thomas PM (2015) ProSight lite: graphical software to analyze top-down mass spectrometry data. *Proteomics* 15(7):1235–1238. doi:[10.1002/pmic.201570050](https://doi.org/10.1002/pmic.201570050)
 18. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2005) The universal protein resource (UniProt). *Nucleic Acids Res* 33(Database issue):D154–D159. doi:[10.1093/nar/gki070](https://doi.org/10.1093/nar/gki070)
 19. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol Cell Proteomics* MCP 9(12):2772–2782. doi:[10.1074/mcp.M110.002766](https://doi.org/10.1074/mcp.M110.002766)
 20. Carvalho PC, Xu T, Han X, Cociorva D, Barbosa VC, Yates JR 3rd (2009) YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics* 25(20):2734–2736. doi:[10.1093/bioinformatics/btp489](https://doi.org/10.1093/bioinformatics/btp489)
 21. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24(21):2534–2536. doi:[10.1093/bioinformatics/btn323](https://doi.org/10.1093/bioinformatics/btn323)
 22. Roepstorff P, Fohlman J (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* 11(11):601. doi:[10.1002/bms.1200111109](https://doi.org/10.1002/bms.1200111109)
 23. Meng F, Cargile BJ, Miller LM, Forbes AJ, Johnson JR, Kelleher NL (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat Biotechnol* 19(10):952–957. doi:[10.1038/nbt1001-952](https://doi.org/10.1038/nbt1001-952)

Mapping Biological Networks from Quantitative Data-Independent Acquisition Mass Spectrometry: Data to Knowledge Pipelines

Erin L. Crowgey, Andrea Matlock, Vidya Venkatraman, Justyna Fert-Bober, and Jennifer E. Van Eyk

Abstract

Data-independent acquisition mass spectrometry (DIA-MS) strategies and applications provide unique advantages for qualitative and quantitative proteome probing of a biological sample allowing constant sensitivity and reproducibility across large sample sets. These advantages in LC-MS/MS are being realized in fundamental research laboratories and for clinical research applications. However, the ability to translate high-throughput raw LC-MS/MS proteomic data into biological knowledge is a complex and difficult task requiring the use of many algorithms and tools for which there is no widely accepted standard and best practices are slowly being implemented. Today a single tool or approach inherently fails to capture the full interpretation that proteomics uniquely supplies, including the dynamics of quickly reversible chemically modified states of proteins, irreversible amino acid modifications, signaling truncation events, and, finally, determining the presence of protein from allele-specific transcripts. This chapter highlights key steps and publicly available algorithms required to translate DIA-MS data into knowledge.

Key words Citrullination, Data-independent acquisition, Phosphorylation, Post-translational modifications, Protein networks, SWATH

1 Introduction

The combination of liquid chromatography (LC) and tandem mass spectrometry (MS/MS) is a technology frequently applied to high-throughput peptide and protein identification and quantification. The most common strategy for peptide identification utilizes a data-dependent acquisition (DDA) approach (review; *see* Bantscheff et al. [1]) [2, 3]. In this approach, the instrument sequentially surveys the observable peptide ions that elute from the LC column at a particular time (MS1 scans), followed by consecutive individual ion isolation and fragmentation events that ultimately assay only a subset of peptides based on their signal intensity and limited

by time to generate their MS/MS (MS2) fragmentation spectra. The acquired mass spectra, both the parent ion mass and a subset of corresponding fragmentation spectra, are matched against theoretical spectra (generated from a sequence database) by a search engine, which then assigns peptide sequences and infers the corresponding proteins (review; *see* Nesvizhskii [4]).

DDA allows the identification of proteins in simple to complex samples. However, as sample complexity increases, there can be inconsistent sampling of all the same ions across a sample set due to the semi-stochastic selection of precursor ions and time dependent selection of co-eluting precursor ions. Therefore, DDA discovery proteomic experiments will have some missing data [5], particularly for the lower abundant ions. Some improvement in the coverage of the ions selected for MS/MS is observed by conducting multiple technical replicates. An alternative to DDA-MS workflow is data-independent acquisition (DIA) (Fig. 1). In DIA, MS/MS scans are collected systematically and independent of

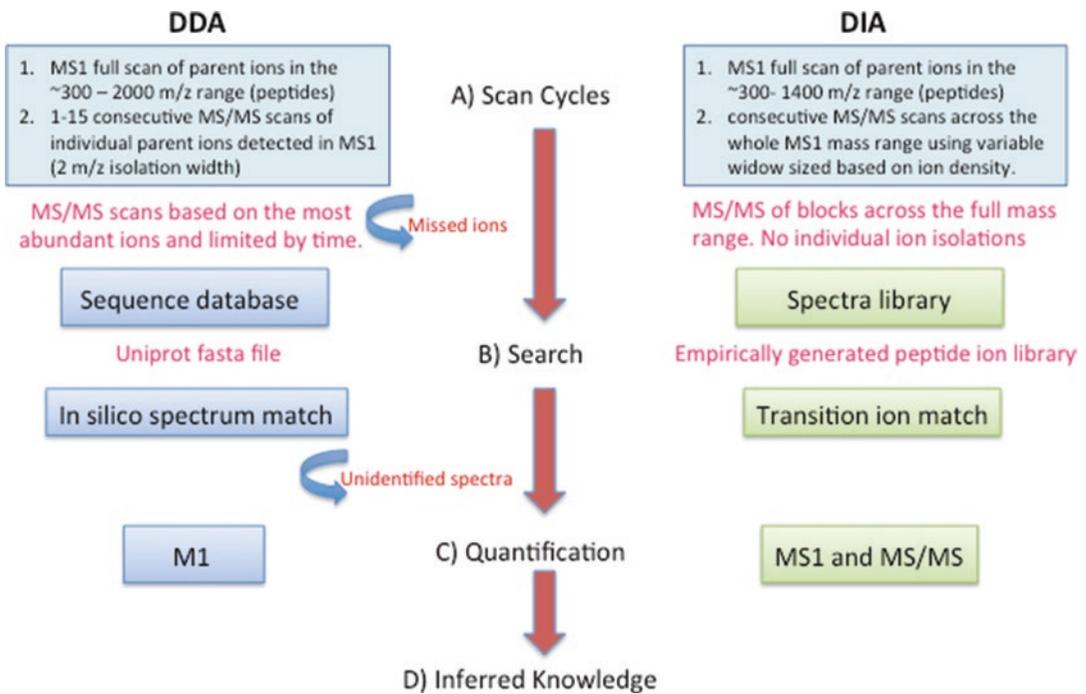


Fig. 1 Data-dependent acquisition mass spectrometry vs. data-independent acquisition mass spectrometry. (A) Scan cycles: DDA, only fragment ion (MS/MS) spectra for selected precursor ions detectable in a survey (MS1) scan are generated. DIA, fragment ion spectra (MS/MS) for all the analytes detectable within the m/z precursor range are recorded. (B) Search: DDA, fragment ion spectra are assigned to their corresponding peptide sequences by sequence database searching. DIA analysis are based on targeted data extraction, in which peptide ions from a spectral library are queried against experimental data to find the best matching fragment ion masses and respective intensities. (C) Quantification: DDA, peptides (and then proteins) are quantified using MS1 signal or spectral counts. DIA computes protein abundance based on selection of transition ion from MS/MS spectra. (D) Translation of large-scale peptide/proteins quantified into knowledge

precursor information over a number of preselected mass ranges. The DIA strategies are based on acquiring fragment ion information for “all” observable precursor ions by repeatedly cycling through predefined sequential m/z windows (DIA-MS/MS spectra) normally over the complete chromatographic elution range generating mixed fragment ion spectrum of all analytes that exist in the m/z range covered (*see Note 1*). We will concentrate on sequential window acquisition of all theoretical (SWATH) mass spectra, which is a two-step process that combines DIA with a preselected peptide library that is used for quantification [6, 7]. This DIA-MS approach is an extension of other MS approaches such as (1) the collection of fragmentation data without precursor ion selection [8], (2) the use of ion mobility spectrometry-CID-time of flight mass spectrometry [9], (3) the use of wide isolation windows [10], and (4) the use of narrow isolation windows combined with many injections [11].

The key advantage of DIA is the ability to reproducibly measure same preset proteins across multiple samples. By design, a DIA experiment ensures selection of the observable ions present for a particular MS/MS analysis and thereby increases the chance of including proteotypic peptides, i.e., the peptides experimentally proven to uniquely identify each protein and be consistently observed for a given protein. A goal should be that the characteristics of these selected peptides attempt to match those used for high-quality targeted MS-based approaches like those of multiple reaction monitoring (MRM) assays [12]. If possible, these preselected proteotypic peptides that represent the targeted proteins within the library would allow accurate protein quantification and infer a lower limit of detection. Furthermore, because DIA provides fragmentation information for each ion within the selected mass range, the analysis will include peptides containing post-translationally modified PTM amino acid residue(s), specific splice variants, as well as any peptides carrying a non-synonymous single-nucleotide polymorphism (SNP).

Many of the DIA-MS studies to date have focused on quantifying proteins within a biological sample and have not included the analysis of site-specific PTMs. However, our premise is that PTMs are potential disease-specific markers of mis-regulated mechanism/pathways. Their importance to biological interpretation is irrefutable, and a large-scale in-depth understanding of protein PTMs is important for gaining a perception of a wide array of cellular functions. The computational analysis of modified peptides was pioneered 20 years ago by Yates et al. [13] and Mann and Wilm [14] and is still an active area of research. Currently, there are >300 types of PTMs described [15, 16], and most likely there are many more yet to be discovered. PTM-search algorithms can be categorized into three groups: targeted, untargeted, and de novo PTM-search methods [17].

DIA-MS techniques overcome the scalability that limits targeted MRM assays to a short list of preselected peptides. In practice, all peptides within the defined mass-to-charge (m/z) window are fragmented collectively in m/z blocks across the full m/z range being covered in DIA [7]. DIA-MS experiments are capable of producing quantitative data on many proteins within a single DIA run but require a library against which to search. Thus, often this results in a shorter time frame per sample analysis than traditional DDA shotgun experiments, which can rely on multidimensional fractionation or long MS run time to obtain the same depth [6, 7]. However, the development of a new library can be time-consuming, and many aspects are still being determined about what constitutes an appropriate library. The most popular DIA-MS platforms currently are Q Exactive Orbitrap mass spectrometer from Thermo Fisher Scientific (Orbitrap), the SCIEX SWATH 2.0, together with their recent triple TOF system (TripleTOF 6600™), followed by high-definition MSE (HDMSE) and ultra-definition MSE (UDMSE) by Waters.

PTMs are often sub-stoichiometric and therefore the less abundant species of molecules within any proteome sample. The gain in sensitivity comes at the expense of the wider isolation windows that may interfere with the ability to accurately confirm PTM localization. This is not a problem if there are not multiple residues within the peptide that could possess the PTM or if the exact residue is not of interest. Otherwise, a secondary experiment for proper confirmation maybe required. Additionally, DIA-MS experiments can provide an alternative approach to PTM identification that depends upon established peptide transition ions from a pre-generated ion library rather than matching an *in silico* digest of specified PTMs. All modified versions of a given peptide with shared transition ions of the unmodified fragment ions allowing one to identify novel PTMs of a given peptide based on parent ion mass shifts and changes in retention time, if expected. This will open additional opportunities to discover (and potentially quantify) unanticipated modified peptide species from DIA data sets by a strategy that does not suffer from the combinatorial explosion of the search space usually experienced with traditional PTM database search approaches.

As mentioned above, PTM analysis by DIA-MS has proven to be extremely useful in the analysis and quantitation of citrullinated proteins. Citrullination is an irreversible deamidation of arginine residues within a protein carried out by enzymatic reaction. This modification leads to the loss of a positive charge and reduction in hydrogen-bonding ability [18]. This modification plays a role in several physiological and pathological processes such as epigenetics, apoptosis, and cancer. However, it is rarely studied because a citrullinated protein or peptide is difficult to discern from its native non-citrullinated form and because the PTM is low abundant, necessitating highly specific and sensitive detection techniques.

A recent publication by Fert-Bober showed how building tissue-specific PTM library increased the accurate detection and quantification of low abundant citrullinated peptides that would have not been possible otherwise [19].

There are a number of features that MS data search algorithms share with respect to preprocessing and post-processing MS data, although the method, format, and information provided can vary significantly. Common features include the handling of protein and peptide sequences, the parsing of results from various proteomics search engine output files, the visualization of MS-related information, and the inference of biological interpretation. Robust tools for data analysis are required to analyze the MS/MS spectra and to translate these large-scale proteome data into biological knowledge. In the area of DIA informatics, there are several computational software/algorithms available for analyzing DIA data (Table 1) that perform the extraction of peptide identifications and quantitation from the raw spectral data files using an empirically generated spectra library, which can be derived from DDA [20] or DIA data. We will not discuss the pros and cons of these in this chapter. However, this chapter will provide an overview of commercially available bioinformatics tools, with the primary focus on the open-source algorithms that researchers can employ when converting DIA-MS data into knowledge.

2 Materials

1. Access to the internet.
2. Processed list of proteins and PTMs identified and quantitated in a DIA-MS experiment (e.g., *see* [19]). See Table 1 for a list of commonly used DIA-MS algorithms.
3. Walkthrough examples include an example output from open-source software Cytoscape: <http://www.cytoscape.org/> and App Store: <http://apps.cytoscape.org/>.

3 Methods

1. *Library considerations*: Each LC-MS run generates data consisting of peak intensities for 1000s of peptides each with specific retention time (RT) and mass-to-charge ratio (m/z) values. These aspects have been recently reviewed [25] and are common to both DIA and DDA. However, there are several issues such as building of MS peptide libraries and LC alignment (e.g., the use of exogenous and endogenous retention time standards) that are unique to DIA, although new methods are being developed and some do not need external retention time calibration

Table 1
Common software applications for the analysis of DIA-MS spectra data

Software	Input spectra format	Type of quantitation	Reference	Website
Skyline	mzML, mzXML, MS2 mz5, and other vendor-specific formats	MS2	MacLean et al. 2010 [22]	https://brendanx-uw1.gs.washington.edu/labkey/project/home/software/Skyline/begin.view
OpenSWATH	mzML, mzXML	MS2	Röst et al. 2014 [7]	http://www.openswath.org
Spectronaut (Biognosys)	HTRMS, WIFF, RAW	MS1, MS2	Reiter et al. 2011 [23]	https://shop.biognosys.ch/spectronaut
PeakView	WIFF	MS2	SCIEX	http://sciox.com/products/software/peakview-software
SWATHProphet	mzML, mzXML	MS2	Keller et al. 2016 [36]	http://tools.proteomecenter.org/wiki/index.php p?title=Software:SWATHProp het
DIA-Umpire	mzXML	MS2	Tsou et al. 2015 [24]	http://diaumpire.sourceforge.net/
Pinnacle	RAW	MS2	http://www.optystech.com/ optystech.com/ home	http://www.optystech.com/

peptides (e.g., DIA-Umpire [24]) as they rely on the retention times of known commonly identified landmark peptides to perform retention time alignment across all the runs.

There are publicly available libraries (e.g., pan-human library) [25], which can be used. But, at the same time, any premade libraries may not contain cell-, organ-, or disease-specific proteins or modified forms that are present in the sample(s) being analyzed. For example, the pan-human library will miss specific stem cell or cardiac proteins that are expressed in specific tissues or cells. The best way to create a library is not entirely clear, and sample type and experimental goals should be considered when addressing this consideration. To date, we have primarily created sample-specific libraries. We hope to move toward increasing our specificity in the selection of peptides and transitions and to perform in DIA with tight percent coefficient of variance in a manner similar to the development of MRM- or SRM-targeted protein assays (*see Note 2*). However, additional work by the scientific community remains to determine how to efficiently generate these peptide libraries.

PeakView (<http://sciex.com/products/software/peakview-software>) and OpenSWATH [7] (<http://www.openswath.org/>), two commonly used algorithms for DIA-MS analysis, both rely on RT alignment of specified ions or a set of standard ions across the elution profile from any DIA run to the peptide library used to make proper peptide identifications. Peptide assignments are initially based on their parent mass, retention time (RT), a set of their fragment ions, and the ratios of fragment ion-relative abundances. Once peptide identification has been performed, quantitation is determined using a set of peptide fragment ions.

PeakView is available through SCIEX (<http://sciex.com/products/software/peakview-software>) and is a stand-alone software application that is compatible with all SCIEX mass spectrometer systems for the quantitative review of LC-MS and MS/MS data. For detailed methods from our group, see [26]. OpenSWATH [7] is available for download by the ETH Zurich group, http://www.openswath.org/openswath_instructions.html, and they provide a well-documented tutorial for executing these processes. Both algorithms, PeakView and OpenSWATH, depend on a library for the DIA analysis.

2. *Protein quantification*: In order to compare LC-MS protein quantification between samples (technical or biological), we recommend, when possible, that they are analyzed simultaneously. We have found this to reduce the extent of normalization required. This is a challenging task as (1) variation in exogenous-supplied (iRT) [27] or endogenous (cIRT) [28] RT standards can exist across multiple runs due to the LC instrument conditions, (2) variability in sample load and the

complexity of peptide mixtures, (3) variation in m/z values due to occasional drift in the calibration of the mass spectrometry instrument, and (4) variation in peak intensities due to spray conditions (in most cases, this is proportional to concentration of peptides in the sample). Thus, alignment with respect to m/z and RT is necessary for quantitative comparison of proteins/peptides (*see Note 3*).

There are several methods for quantifying protein concentrations from DIA-MS data, including MSstats and OpenSWATH, and have been recently reviewed [20]. MSstats (<https://www.bioconductor.org/packages/release/bioc/html/MSstats.html>), an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments [29], is an R/Bioconductor package. It provides protein abundance using linear-mixed model and group comparisons. Differential analysis can be helpful when comparing a disease/perturbation to a control/background.

3. *Data to knowledge – overlay on interactome*: Proteomic discovery-based data ultimately shifts the burden to the downstream analysis, which requires an extensive systems biology approach for data interpretation. A researcher is dealing with the quantifiable proteome, or interactome depending on the experiment and questions being assessed.

A major advantage to using UniProtKB/Swiss-Prot accessions (and sequences) for protein identification is the ability to link protein functional annotations to the proteins identified within the original library. There are several mapping tools available to convert between various database identifiers, for example, the UniProt mapping tool (<http://www.uniprot.org/uploadlists/>) can map UniProtKB accession numbers to other database identifiers such as RefSeq. These mapping steps are often required for downstream analysis and ultimately allow the connection of many underlying functional databases.

STRING (<http://string-db.org/>) [30] is a database of known and predicted protein interactions, which include direct and indirect associations that are derived from five sources: genomic context, high-throughput experiments, co-expression, automated text mining, and previous knowledge in databases. It has a user-friendly interface and provides flexibility for setting various parameters like confidence. The output from STRING can also be saved in a text file format compatible with Cytoscape and therefore allows further customization for visualizing the results.

NetworkAnalyst (<http://www.networkanalyst.ca/>) is another open-source tool available for analyzing a list of proteins using protein-protein interaction networks [31, 32]. It has an easy-to-use graphical interface, and proteins of interest are mapped to manually curated protein-protein interaction

database to construct relevant networks. The tool provides a system for functional enrichment analysis and the results, including the network map, can be easily exported.

Cytoscape, an open-source bioinformatics software for visualizing molecular interaction networks, can be freely downloaded from <http://www.cytoscape.org> (Fig. 2). Furthermore, Cytoscape has several plug-ins (applications) (<http://apps.cytoscape.org/apps/all>) available for creating an interactome from a user input (i.e., lists of proteins). The apps can easily be installed by using the Application Manager within Cytoscape (Fig. 2). Reactome FI, which was originally developed for microarray data, allows a researcher to upload a list of genes of interest, and the algorithm will display the interactome for those candidates. There are many types of analysis that Reactome FI can execute, and the user can easily select the appropriate module (Fig. 3).

Reactome FI enables functional enrichment analysis and easily customizable network output maps. Using the DIA-MS

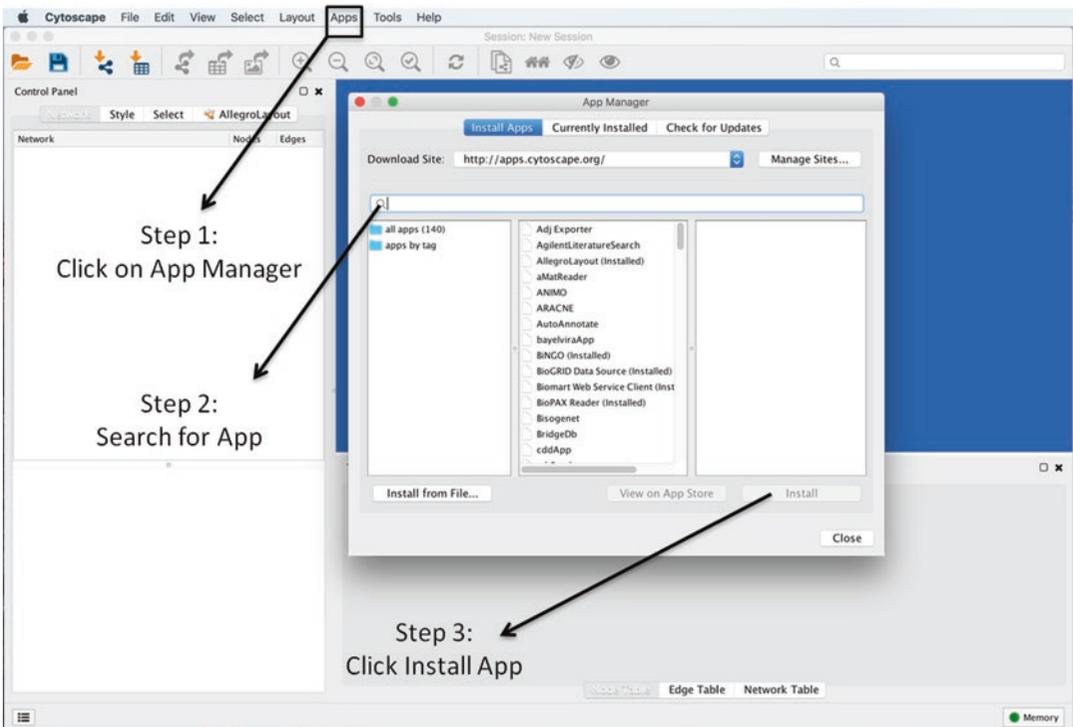


Fig. 2 Download Cytoscape and install plug-ins. (1) Cytoscape is an open-source application that can be downloaded from <http://www.cytoscape.org>. The application is free and available for Mac or Windows. The application requires Java, which is also freely available at <http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html>. (2) Install plug-ins. *Step 1:* Open Cytoscape and click on the Apps tab. The App Manager will appear. *Step 2:* Type in the search bar the name of the application of interest. *Step 3:* Select the application of interest and click on install

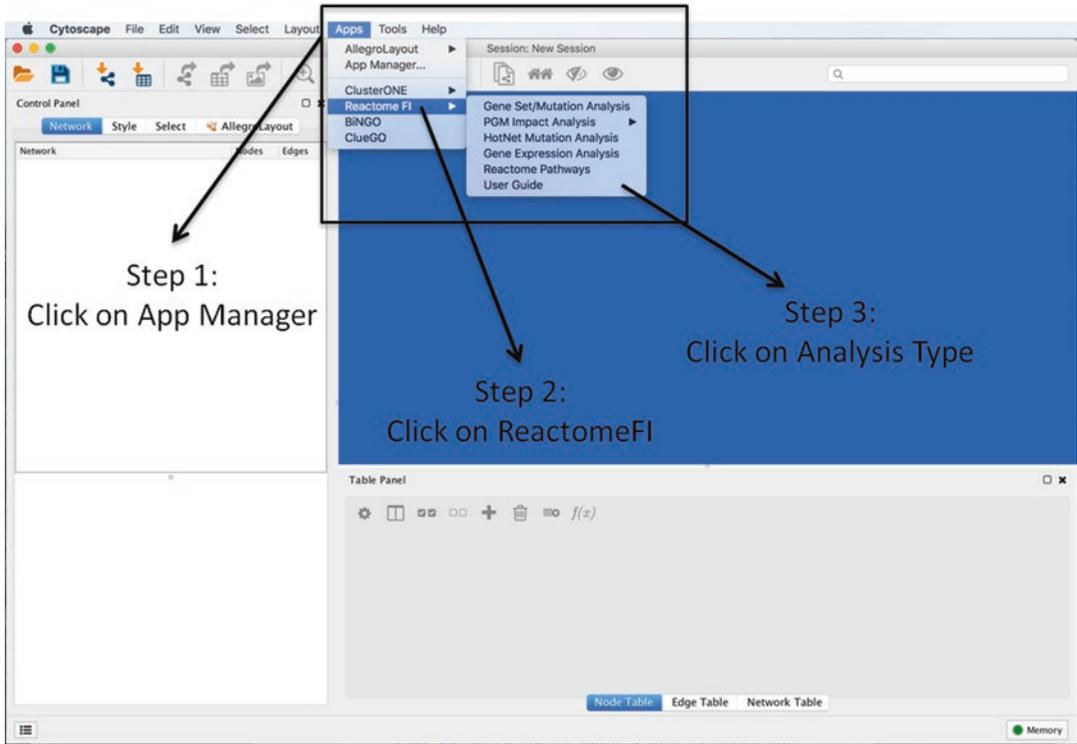


Fig. 3 Overview of Reactome FI plug-in. *Step 1:* Click on the apps tab and application installed (following Fig. 2) will appear. *Step 2:* Select the application of interest (i.e., Reactome FI). *Step 3:* Select the type of analysis to execute

(SWATH) data released in [19], an interactome was generated for citrullinated proteins that were differentially regulated in cardiocytes [19] (Fig. 4). The results from Reactome FI can also be analyzed via other Cytoscape plug-ins, including ClusterONE [33] (Fig. 5).

4. *Data to knowledge – functional enrichment analysis.* Enrichment analyses are typically performed utilizing gene ontology (GO) annotations, which include cellular compartment, molecular function, and biological process, and/or pathway annotations from databases like KEGG and Reactome. Common statistical tests include Fisher exact or a p -value probability or chance of seeing at least x number of genes out of the total n genes in the list annotated to a particular GO. The goal is to differentiate enrichment above random background. These types of analysis can help generate new hypothesis about protein dynamics under a given biological system or enhance our current understanding of biological processes associated with a given condition.

There are several publicly available tools for enrichment analysis, including iProXpress (<http://pir.georgetown.edu/iproXpress2/>), BiNGO [34], and Panther (<http://geneontol->



Fig. 4 Reactome FI analysis of the top citrullinated proteins in heart diseases. The top citrullinated proteins from [19] were uploaded and analyzed in Cytoscape via Reactome FI. *Circles nodes* represent proteins that were differentially citrullinated, whereas *triangle nodes* represent proteins that were not reported as having differentially citrullinated residues, but are linked to proteins, through protein-protein interactions (*gray lines*) that do have differentially regulated citrullinated residues. The top three pathways enriched per module were extracted. *Module 1*: Striated muscle contraction, hypertrophic cardiomyopathy, and dilated cardiomyopathy. *Module 2*: glycolysis/gluconeogenesis. Biosynthesis of amino acids and validated targets of c-Myc transcriptional activation. *Module 3*: Parkinson's disease, the citric acid cycle and respiratory electron transport, and Huntington's disease. *Module 4*: Tyrosine metabolism, fatty acid degradation, and retinol metabolism. *Module 5*: The citric acid (TCA) cycle and respiratory electron transport, carbon metabolism, and metabolic pathway

ogp.org/page/go-enrichment-analysis). BiNGO [34] is a Cytoscape plug-in (*see* Fig. 2 for overview on downloading plug-ins) that enables functional enrichment analysis and visualization. Data can be uploaded directly from the user or from a network, and the user has the option to select various parameters and statistical tests for analyzing the data (Fig. 6). Using the DIA-MS (SWATH) data released in a recent publication, an example of a functional enrichment output from BiNGO [34] was generated for citrullinated proteins that were differentially regulated in cardiocytes [19] (Fig. 7).

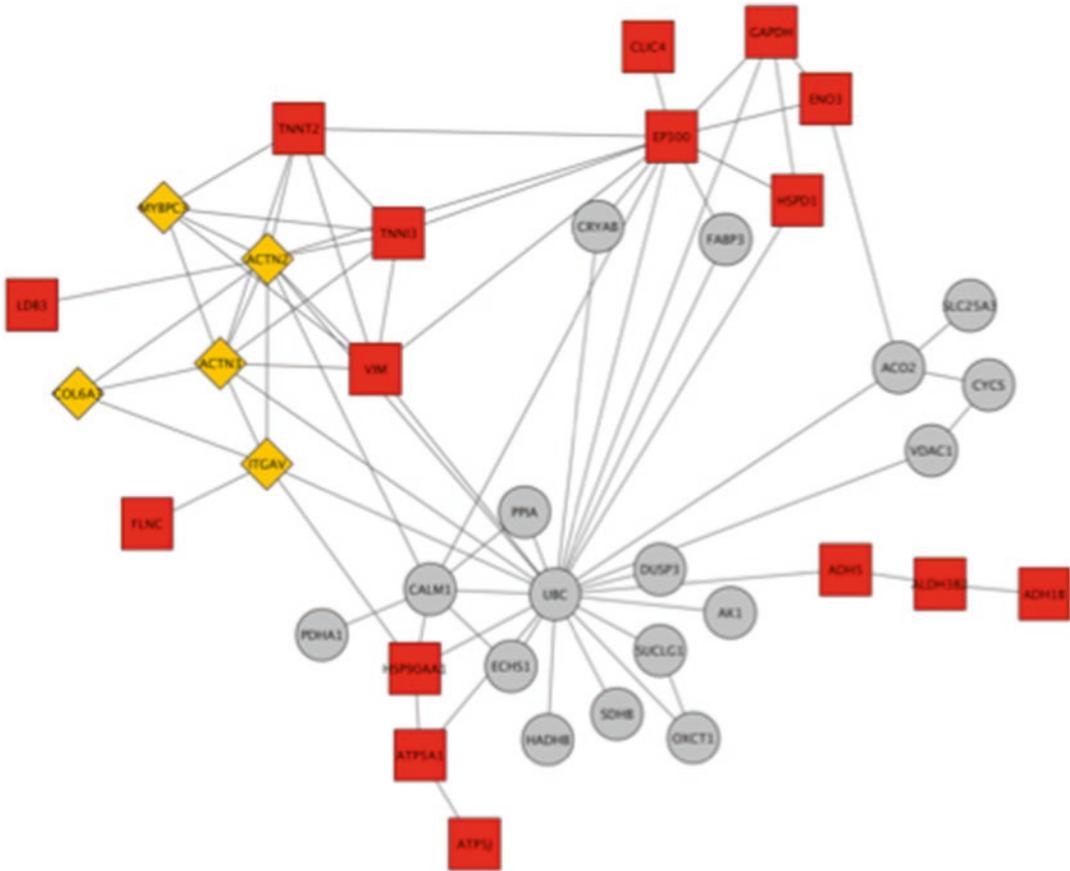


Fig. 5 ClusterONE analysis of an interactome. The network in Fig. 4 was further analyzed in Cytoscape using ClusterONE. *Orange triangles* are nodes that represent proteins that are highly connected within and across modules. *Red squares* are nodes that were clustered, whereas *gray circles* are outliers. The top cluster consisted of eight proteins: COL6A3, ACTN2, ACTN3, ITGAV, VIM, TNNT2, MYBPC3, and TNNI3 (p -value 0.001, density 0.714, quality 0.625)

5. *Considerations for PTM analysis:* Diversity at the protein level comes from (1) mRNA splice variants and internal start sites, (2) variants affecting the primary sequence of amino acids (e.g., SNPs), and (3) different PTMs. The large-scale study of variance due to genomic alterations typically requires next-generation sequencing techniques for DNA/RNA molecules, as a priori knowledge is not required. Proteomic data is further complicated as different forms of PTMs may occur in tandem, greatly increasing the complexity of the proteome. PTMs broadly contribute to the recent explosion of proteomic data as they possess a significant aspect to protein function.

In PTM analysis, each peptide representing a modification site of interest needs to stand alone; this is in contrast to proteome analysis where several peptides are usually taken into consideration to reveal characteristics of a single protein. Global

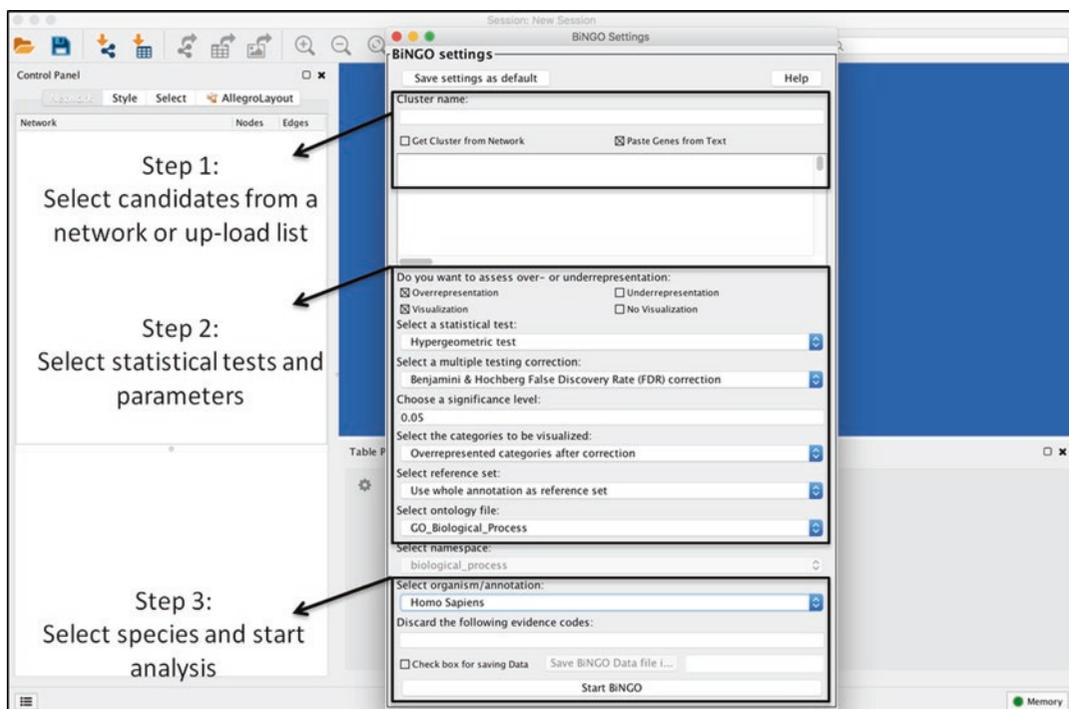


Fig. 6 Overview for executing a BiNGO analysis. *Step 1*: Enter the name of the analysis, and select either “Get Cluster from Network” or “Paste Genes from Text.” *Step 2*: Select over- or underrepresentation; select a statistical test (i.e., hypergeometric); and select a multiple testing correction (i.e., Benjamini and Hochberg false discovery rate (FDR) correction), a significance level, the categories to be visualized, reference set, and ontology type (biological process, molecular function, or cellular compartment). *Step 3*: Select the appropriate species and start BiNGO analysis

PTM analysis remains a major challenge in the field and is very resource demanding. MS-based proteomics provides tens of thousands of sites, raising the question of their biological relevance. Researchers are faced with the challenge of selecting biologically relevant sites from large-scale data, and then determining how to perform functional follow-up assays on these candidates (*see Note 4*).

Oftentimes, it can be challenging to determine the exact location of a PTM, such as phosphorylation, and as such, there are several publicly available algorithms that can be applied to further assess PTM data generated from MS/MS data. For example, Ascore is an algorithm that measures the probability of a correct phosphorylation site localization based on the presence and intensity of site-determining ions in MS/MS spectra [35]. Recently, PTMProphet was added to the SWATHProphet software and serves as a tool to identify/annotate modifications in peptide sequences by identifying precursor ions consistent with a modification, along with the mass and localization of the modification in the peptide sequence [36]. Another algorithm for scoring/annotating

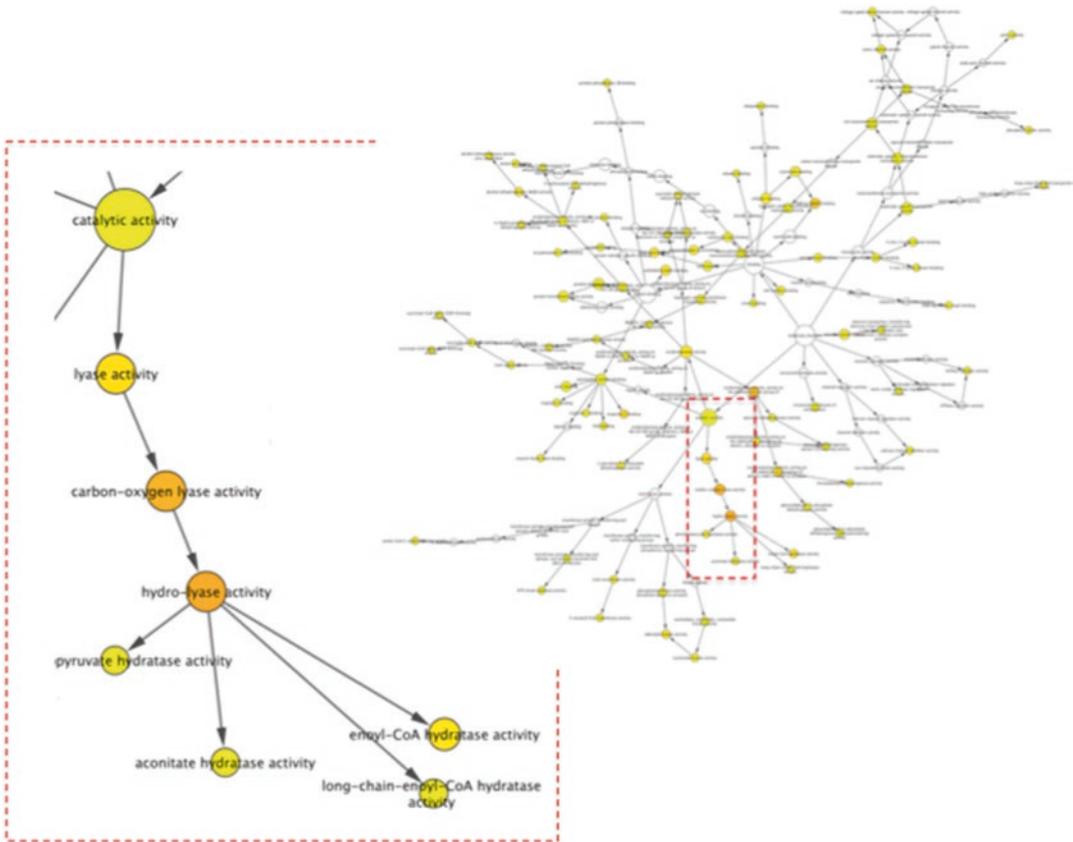


Fig. 7 Gene ontology (molecular function) enrichment analysis using BiNGO. The top citrullination proteins from [19] were uploaded into BiNGO and analyzed for enriched gene ontology (GO) molecular function terms. *Orange nodes* represent the most significantly enriched GO terms, whereas *white* and *yellow* represent the least significantly enriched GO terms. The BiNGO analysis highlights the hierarchy of the ontological terms. For this dataset, the most enriched molecular function terms were hydro-lyase activity, carbon-oxygen lyase activity, oxidoreductase activity, NAD or NADH binding, lyase activity, troponin C binding, and enoyl-CoA hydratase activity

PTM localization, called LuciPHOr, uses a modified target-decoy-based approach that uses mass accuracy and peak intensities for site localization scoring and false localization rate (FLR) [37].

Citrullination is another important PTM and recent advancements have enabled DIA-MS detection, and bioinformatics methods have enabled the analysis for this biologically relevant irreversible PTM, including algorithms for scoring the location of this PTM within a peptide sequence [19]. Interestingly, it has been supported that citrullination of sarcomere proteins causes a decrease in $\text{Ca}^{(+2)}$ sensitivity in skinned cardiomyocytes, indicating an important structural and functional alteration associated with this PTM [19].

6. *Considerations for translating large proteome datasets into biological knowledge leveraging PTM data:* For DIA-MS data, it is

important at both the peptide level and protein level to translate the data into an integrated knowledge base. Translating large data into knowledge is a difficult task, and there is no gold standard or process available. At the foundation, a system capable of linking many functional annotations together is essential. Without this type of connectivity, functional enrichment analysis such as GO and pathway would not be possible. Linking to disease databases like the Online Mendelian Inheritance in Man (OMIM), a catalog of human Mendelian disorders, which contains 20,267 entries describing 13,606 genes from ~7000 disorders [38], helps to provide a resource for annotating clinically relevant gene/protein candidates.

Text and data mining also become invaluable resources when analyzing large datasets as it is impossible to manually research all quantified proteins in their various states across many experiments. Data mining involves integration of many biological datasets and annotations and when utilized effectively can produce more holistic insights. See review [39] for a comparison of different methods for data integration and their advantages and disadvantages. Furthermore, there are other tools for data integration and text mining, including RapidMiner, KNIME, and R statistics, that can be customized for the end users' goals.

Oftentimes, a researcher may need to leverage ortholog mapping across species to reveal significant findings. This underutilized aspect holds considerable value and can infer importance of a modifiable amino acid residue. This is particularly important because functionally important modification sites are more likely to be evolutionarily conserved; yet cross-species comparison of PTMs is difficult since they often lie in structurally disordered protein domains. Current tools that address this are PhosphoSitePlus [40], Phospho.ELM [41], Phosphorylation Site Database [42], PHOSIDA [43], PhosphoAt [44], PhosphOrtholog [45], NetworKIN [46], and RegPhos [47].

O-GLYCBASE [48] and dbOGAP [49] are databases for glycoproteins, most of which include experimentally verified O-linked glycosylation sites. UbiProt [50] stores experimental ubiquitylated proteins and ubiquitylation sites, which are implicated in protein degradation through an intracellular ATP-dependent proteolytic system. Furthermore, PTMScout (<http://ptmscout.mit.edu>) is another web resource that is constructed around a custom database of PTM experiments and contains information from external protein and post-translational resources, including GO annotations, Pfam domains, and Scansite predictions of kinase and phosphopeptide-binding domain interactions [51].

Motif analysis strategies and domain-domain interactions related to PTMs are also important aspects in translating data. Proteins having related functions may not show overall high

sequence similarity, yet they may contain sequences of amino acid residues that are highly conserved within the tertiary structure of the protein. Currently, the largest collection of sequence motifs in the world is PROSITE [52] and META Site such as MOTIF [53]. PROSITE can be accessed via ExPASy (<http://www.expasy.org>). A free software package named MacPattern [54] is available for searching PROSITE motifs. Other useful resources for searching protein motifs are BLOCKS [55], MOTIF Search (<http://www.genome.jp/tools/motif/>), and MoST [56].

SysPTM [57] has designed a systematic platform for multi-type PTM research and data mining. Additionally, Human Protein Reference Database (HPRD) [58] contains a wealth of information relevant to the function of human proteins in health and disease, as well as the annotation of PTMs.

The rate of discovery for PTMs is gaining momentum and is significantly outpacing our biological understanding of the function and regulation of these modifications, and data mining techniques have the potential to enable the discovery of previously unknown patterns and relationships hidden in large datasets.

4 Notes

1. DIA methods have advantages and disadvantages related to the instrument and the composition and complexity of the biological sample. Similar to DDA experiments, the instrument method of DIA experiments entails trade-offs across mass resolution and mass accuracy, scanning rates and the number of data points taken across a peak, number and width of isolation windows, and cycle times. Ultimately, DIA encompasses the strengths of both DDA and MRM approaches [6], combining shotgun (DDA) discovery proteomics with the quantitative capabilities and high-throughput nature of targeted approaches [7].
2. A spectral library of identified peptides can be manually programmed, downloaded (if available), or generated by previous DDA experiments. The effectiveness of sequence-searching approach depends on (a) high-quality reference spectra, with good signal-to-noise ratios and devoid of impurities, and (b) effective matching algorithms with the robustness and flexibility to accommodate imperfect matches while minimizing false matches.
3. It is our experience that the larger the number of samples being compared, complexity of data analysis increases due to limited scalability of current methods. It is also our experience that only selected peptides and transitions with coefficient of variances of under 20 % in the DDA runs are necessary to ensure

accurate quantification. This is the level of reproducibility and precision required in clinical chemistry hospital assays.

4. The interpretation of proteome data obtained from high-throughput methods cannot be appropriately deciphered without a priori knowledge which may come by biochemical or physiological data where specific PTM data from in vitro or in vivo is available. For example, cardiac troponin (cTnI) plays a key role in the regulation of contraction and relaxation of heart muscle. There are numerous phosphorylation sites on cTnI with and without in vitro or in vivo PKA phosphorylation [59]. Mutational data supports that residues that have been substituted as a pseudo mimetic, such as the phosphorylation of sites 22 and 23 in cTnI being replaced with Asp to mimic the negative charge, have a profound effect on the function of cTnI [60].

References

1. Bantscheff M, Lemeer S, Savitski MM, Kuster B (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* 404:939–965. doi:[10.1007/s00216-012-6203-4](https://doi.org/10.1007/s00216-012-6203-4)
2. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422:198–207. doi:[10.1038/nature01511](https://doi.org/10.1038/nature01511)
3. Zhang Y, Fonslow BR, Shan B et al (2013) Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* 113:2343–2394. doi:[10.1021/cr3003533](https://doi.org/10.1021/cr3003533)
4. Nesvizhskii AI (2007) Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol* 367:87–119. doi:[10.1385/1-59745-275-0:87](https://doi.org/10.1385/1-59745-275-0:87)
5. Bateman NW, Goulding SP, Shulman NJ et al (2014) Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Mol Cell Proteomics* 13:329–338. doi:[10.1074/mcp.M112.026500](https://doi.org/10.1074/mcp.M112.026500)
6. Gillet LC, Navarro P, Tate S et al (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* 11:O111.016717. doi:[10.1074/mcp.O111.016717](https://doi.org/10.1074/mcp.O111.016717)
7. Röst HL, Rosenberger G, Navarro P et al (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 32:219–223. doi:[10.1038/nbt.2841](https://doi.org/10.1038/nbt.2841)
8. Purvine S, Eppel J-T, Yi EC, Goodlett DR (2003) Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* 3:847–850. doi:[10.1002/pmic.200300362](https://doi.org/10.1002/pmic.200300362)
9. Myung S, Lee YJ, Moon MH et al (2003) Development of high-sensitivity ion trap ion mobility spectrometry time-of-flight techniques: a high-throughput nano-LC-IMS-TOF separation of peptides arising from a *Drosophila* protein extract. *Anal Chem* 75:5137–5145. doi:[10.1021/ac030107f](https://doi.org/10.1021/ac030107f)
10. Venable JD, Dong M-Q, Wohlschlegel J et al (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* 1:39–45. doi:[10.1038/nmeth705](https://doi.org/10.1038/nmeth705)
11. Panchaud A, Jung S, Shaffer SA et al (2011) Faster, quantitative, and accurate precursor acquisition independent from ion count. *Anal Chem* 83:2250–2257. doi:[10.1021/ac103079q](https://doi.org/10.1021/ac103079q)
12. Carr SA, Abbatiello SE, Ackermann BL et al (2014) Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Mol Cell Proteomics* 13:907–917. doi:[10.1074/mcp.M113.036095](https://doi.org/10.1074/mcp.M113.036095)
13. Yates JR, Eng JK, McCormack AL, Schieltz D (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 67:1426–1436
14. Mann M, Wilm M (1995) Electrospray mass spectrometry for protein characterization. *Trends Biochem Sci* 20:219–224
15. Witze ES, Old WM, Resing KA, Ahn NG (2007) Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* 4:798–806. doi:[10.1038/nmeth1100](https://doi.org/10.1038/nmeth1100)

16. Jensen ON (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol* 8:33–41. doi:[10.1016/j.cbpa.2003.12.009](https://doi.org/10.1016/j.cbpa.2003.12.009)
17. Tsur D, Tanner S, Zandi E et al (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol* 23:1562–1567. doi:[10.1038/nbt1168](https://doi.org/10.1038/nbt1168)
18. György B, Tóth E, Tarcsa E et al (2006) Citrullination: a posttranslational modification in health and disease. *Int J Biochem Cell Biol* 38:1662–1677. doi:[10.1016/j.biocel.2006.03.008](https://doi.org/10.1016/j.biocel.2006.03.008)
19. Fert-Bober J, Giles JT, Holewinski RJ et al (2015) Citrullination of myofibrillar proteins in heart failure. *Cardiovasc Res* 108:232–242. doi:[10.1093/cvr/cvv185](https://doi.org/10.1093/cvr/cvv185)
20. Bilbao A, Varesio E, Luban J et al (2015) Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* 15:964–980. doi:[10.1002/pmic.201400323](https://doi.org/10.1002/pmic.201400323)
21. Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 73:2092–2123. doi:[10.1016/j.jpro.2010.08.009](https://doi.org/10.1016/j.jpro.2010.08.009)
22. MacLean B, Tomazela D, Shulman et al. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26(7):966–968. Doi [10.1093/bioinformatics/btq054](https://doi.org/10.1093/bioinformatics/btq054)
23. Bernhardt OM, Selevsek N, Gillet LC, et al. (2012) Spectronaut A fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. *Proceedings of the 60th ASMS Conference on Mass Spectrometry and Allied Topics, 2012, Vancouver, BC, Canada.*
24. Tsou C-C, Avtonomov D, Larsen B et al (2015) DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* 12:258–264. doi:[10.1038/nmeth.3255](https://doi.org/10.1038/nmeth.3255)
25. Rosenberger G, Koh CC, Guo T et al (2014) A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* 1:140031. doi:[10.1038/sdata.2014.31](https://doi.org/10.1038/sdata.2014.31)
26. Holewinski RJ, Parker SJ, Matlock AD et al. (2016) Methods for SWATH™: data independent acquisition on TripleTOF mass spectrometers. *Methods Mol Biol.* 1410:265–79. doi:[10.1007/978-1-4939-3524-6_16](https://doi.org/10.1007/978-1-4939-3524-6_16).
27. Escher C, Reiter L, MacLean B et al (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* 12:1111–1121. doi:[10.1002/pmic.201100463](https://doi.org/10.1002/pmic.201100463)
28. Parker SJ, Rost H, Rosenberger G et al (2015) Identification of a set of conserved eukaryotic internal retention time standards for data-independent acquisition mass spectrometry. *Mol Cell Proteomics* 14:2800–2813. doi:[10.1074/mcp.O114.042267](https://doi.org/10.1074/mcp.O114.042267)
29. Choi M, Chang C-Y, Clough T et al (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 30:2524–2526. doi:[10.1093/bioinformatics/btu305](https://doi.org/10.1093/bioinformatics/btu305)
30. Szklarczyk D, Franceschini A, Wyder S et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43:D447–D452. doi:[10.1093/nar/gku1003](https://doi.org/10.1093/nar/gku1003)
31. Xia J, Gill EE, Hancock REW (2015) NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat Protoc* 10:823–844. doi:[10.1038/nprot.2015.052](https://doi.org/10.1038/nprot.2015.052)
32. Xia J, Benner MJ, Hancock REW (2014) NetworkAnalyst—integrative approaches for protein-protein interaction network analysis and visual exploration. *Nucleic Acids Res* 42:W167–W174. doi:[10.1093/nar/gku443](https://doi.org/10.1093/nar/gku443)
33. Nepusz T, Yu H, Paccanaro A (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 9:471–472. doi:[10.1038/nmeth.1938](https://doi.org/10.1038/nmeth.1938)
34. Maere S, Heymans K, Kuiper M (2005) BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21:3448–3449. doi:[10.1093/bioinformatics/bti551](https://doi.org/10.1093/bioinformatics/bti551)
35. Beausoleil SA, Villén J, Gerber SA et al (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 24:1285–1292. doi:[10.1038/nbt1240](https://doi.org/10.1038/nbt1240)
36. Keller A, Bader SL, Kusebauch U et al (2016) Opening a SWATH window on posttranslational modifications: automated pursuit of modified peptides. *Mol Cell Proteomics* 15:1151–1163. doi:[10.1074/mcp.M115.054478](https://doi.org/10.1074/mcp.M115.054478)
37. Fermin D, Walmsley SJ, Gingras A-C et al (2013) LuciPHOR: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Mol Cell Proteomics* 12:3409–3419. doi:[10.1074/mcp.M113.028928](https://doi.org/10.1074/mcp.M113.028928)
38. Manwar Hussain MR, Khan A, Ali Mohamoud HS (2014) From genes to health—challenges and opportunities. *Front Pediatr* 2:12. doi:[10.3389/fped.2014.00012](https://doi.org/10.3389/fped.2014.00012)

39. Gligorijević V, Pržulj N (2015) Methods for biological data integration: perspectives and challenges. *J R Soc Interface* 12:20150571. doi:[10.1098/rsif.2015.0571](https://doi.org/10.1098/rsif.2015.0571)
40. Hornbeck PV, Kornhauser JM, Tkachev S et al (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40:D261–D270. doi:[10.1093/nar/gkr1122](https://doi.org/10.1093/nar/gkr1122)
41. Dinkel H, Chica C, Via A et al (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res* 39:D261–D267. doi:[10.1093/nar/gkq1104](https://doi.org/10.1093/nar/gkq1104)
42. Wurgler-Murphy SM, King DM, Kennelly PJ (2004) The phosphorylation site database: a guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms. *Proteomics* 4:1562–1570. doi:[10.1002/pmic.200300711](https://doi.org/10.1002/pmic.200300711)
43. Gnad F, Ren S, Cox J et al (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8:R250. doi:[10.1186/gb-2007-8-11-r250](https://doi.org/10.1186/gb-2007-8-11-r250)
44. Heazlewood JL, Durek P, Hummel J et al (2008) PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res* 36:D1015–D1021. doi:[10.1093/nar/gkm812](https://doi.org/10.1093/nar/gkm812)
45. Chaudhuri R, Sadrieh A, Hoffman NJ et al (2015) PhosphOrtholog: a web-based tool for cross-species mapping of orthologous protein post-translational modifications. *BMC Genomics* 16:617. doi:[10.1186/s12864-015-1820-x](https://doi.org/10.1186/s12864-015-1820-x)
46. Linding R, Jensen LJ, Pasculescu A et al (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res* 36:D695–D699. doi:[10.1093/nar/gkm902](https://doi.org/10.1093/nar/gkm902)
47. Lee T-Y, Bo-Kai Hsu J, Chang W-C, Huang H-D (2011) RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic Acids Res* 39:D777–D787. doi:[10.1093/nar/gkq970](https://doi.org/10.1093/nar/gkq970)
48. Gupta R, Birch H, Rapacki K et al (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 27:370–372
49. Wang J, Torii M, Liu H et al (2011) dbO-GAP—an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics* 12:91. doi:[10.1186/1471-2105-12-91](https://doi.org/10.1186/1471-2105-12-91)
50. Chernorudskiy AL, Garcia A, Eremin EV et al (2007) UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics* 8:126. doi:[10.1186/1471-2105-8-126](https://doi.org/10.1186/1471-2105-8-126)
51. Naegle KM, Gymrek M, Joughin BA et al (2010) PTMScout, a web resource for analysis of high throughput post-translational proteomics studies. *Mol Cell Proteomics* 9:2558–2570. doi:[10.1074/mcp.M110.001206](https://doi.org/10.1074/mcp.M110.001206)
52. Falquet L, Pagni M, Bucher P et al (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res* 30:235–238
53. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res* 30:42–46
54. Fuchs R (1991) MacPattern: protein pattern searching on the Apple Macintosh. *Comput Appl Biosci* 7:105–106
55. Henikoff S, Henikoff JG (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res* 19:6565–6572
56. Tatusov RL, Altschul SF, Koonin EV (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA* 91:12091–12095
57. Li H, Xing X, Ding G et al (2009) SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol Cell Proteomics* 8:1839–1849. doi:[10.1074/mcp.M900030-MCP200](https://doi.org/10.1074/mcp.M900030-MCP200)
58. Keshava Prasad TS, Goel R, Kandasamy K et al (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res* 37:D767–D772. doi:[10.1093/nar/gkn892](https://doi.org/10.1093/nar/gkn892)
59. Zhang P, Kirk JA, Ji W et al (2012) Multiple reaction monitoring to identify site-specific troponin I phosphorylated residues in the failing human heart. *Circulation* 126:1828–1837. doi:[10.1161/CIRCULATIONAHA.112.096388](https://doi.org/10.1161/CIRCULATIONAHA.112.096388)
60. Kooij V, Zhang P, Piersma SR et al (2013) PKC α -specific phosphorylation of the troponin complex in human myocardium: a functional and proteomics analysis. *PLoS One* 8:e74847. doi:[10.1371/journal.pone.0074847](https://doi.org/10.1371/journal.pone.0074847)

Annotation of Alternatively Spliced Proteins and Transcripts with Protein-Folding Algorithms and Isoform-Level Functional Networks

Hongdong Li, Yang Zhang, Yuanfang Guan, Rajasree Menon, and Gilbert S. Omenn

Abstract

Tens of thousands of splice isoforms of proteins have been catalogued as predicted sequences from transcripts in humans and other species. Relatively few have been characterized biochemically or structurally. With the extensive development of protein bioinformatics, the characterization and modeling of isoform features, isoform functions, and isoform-level networks have advanced notably. Here we present applications of the I-TASSER family of algorithms for folding and functional predictions and the IsoFunc, MISOmine, and Hisonet data resources for isoform-level analyses of network and pathway-based functional predictions and protein-protein interactions. Hopefully, predictions and insights from protein bioinformatics will stimulate many experimental validation studies.

Key words Functional prediction, Isoform network, Protein folding, Splice isoforms

1 Introduction

One of the most remarkable developments in biological evolution is the emergence in multicellular organisms of gene structures with exons and introns. An elaborate splicing machinery in cells processes heterogeneous nuclear RNAs and generates several different mRNA transcripts from individual genes that can be translated into protein products. Just describing gene or protein expression as “upregulated” or “downregulated” ignores the fact that these transcripts and proteins are mixtures. These splice variants can and often do have dramatically different functions; when the proteins fold similarly and compete for target sites, they may, in fact, have opposing actions, such as proapoptotic and antiapoptotic activities [1].

Alternative splicing generates protein diversity without increasing genome size. This phenomenon seems to explain how humans can “get by” with only 20,000 protein-coding genes, whereas there

were predictions of 50,000 to 100,000 or more protein-coding genes when the Human Genome Project was launched. The splice variants cannot be identified in genome sequences, but the splicing can be mapped to the gene exon/intron structures. There are multiple kinds of splicing events, including alternative 5' or 3' start sites, mutually exclusive exons (exon swaps), intron retention, alternative promoters, and alternative polyadenylation. There are examples of every kind of splicing in cancers, for example, and complex combinations of splice isoforms are well described in the nervous system. Ensembl, UniProt, neXtProt, RefSeq, and ECGene are databases with extensive information about protein splice variants.

2 Materials

Depending on the biological or clinical question being investigated, either primary experimental data or publicly available datasets for protein and transcript isoforms from appropriate specimens may be utilized for annotation and characterization of splice isoforms. For example, we have characterized splice isoforms of Her2/ERBB2+ breast cancers from humans and in mouse models [2, 3].

1. Use the current version of UniProt (e.g., release 2015/03) at <http://www.uniprot.org> to obtain a reliable, high-quality set of protein isoforms which are consistently annotated both in Ensembl (version 75) at http://wwwensembl.org/Homo_sapiens/Info/Index and in NCBI RefSeq (e.g., release 70) at <http://www.ncbi.nlm.nih.gov/refseq/>.

Note that the annotation on splice isoforms varies from database to database and version to version. In Ensembl, information on protein-coding transcripts for a gene is updated or changed whenever the database version is changed. RefSeq is generally less inclusive than Ensembl.

The identification of a “canonical isoform” often defaults to the longest product (protein sequence). UniProt curators choose a canonical variant from among several protein isoforms encoded by one gene using some mixture of the following criteria: highest expressed (varies by tissue and conditions); most conserved across species; the amino acid sequence that allows the clearest description of domains, isoforms, polymorphisms, and post-translational modifications; or, finally, the longest sequence. The other sequences are called “noncanonical” isoforms: see <http://www.uniprot.org/help/canonicalandisoforms>. As described below, we propose a method to identify the “most highly connected isoform” and consider the highest connected isoform (HCI) the canonical form.

Generally, the functional annotation of genes is based on their widely studied canonical protein or, more crudely, on the mixture of unrecognized isoforms. There is only very limited

experimental evidence or computational annotation of functions of noncanonical protein isoforms. Gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) are widely used; we recommend also using GeneCards for gene-level annotations (<http://www.genecards.org/>).

2. RNA-Seq provides far more precise measurement of levels of expression of transcripts and their isoforms than microarray analyses do. Available datasets can be downloaded from the NCBI Short Read Archive (<http://www.ncbi.nlm.nih.gov/sra>). Often there are biological replicates available.

Analyze the RNA-Seq data using Sailfish [4], an alignment-free algorithm (unlike Tophat/Cufflinks) for estimation of the isoform abundances. Sailfish builds a unique index of all k-mers (short and consecutive sequences containing k nucleic acids), counts the occurrences of the k-mers in the RNA-Seq fragments, and quantitates the transcripts by the number of occurrences of the k-mers through an EM algorithm. The index file for the Sailfish quantitation process can be generated from Ensembl cDNA file (GRCh38 version).

Use the R package edgeR to quantitate differential expression of transcript isoforms between two tissues or tumor types being compared. This method uses an over-dispersed Poisson model to account for biological and technical variability. Bonferroni correction of p values for multiple hypothesis testing can be performed with the `p.adjust` function in stats, R package (<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/p.adjust.html>). Use the Sailfish estimates of read counts for each transcript isoform in calculations of differential expression.

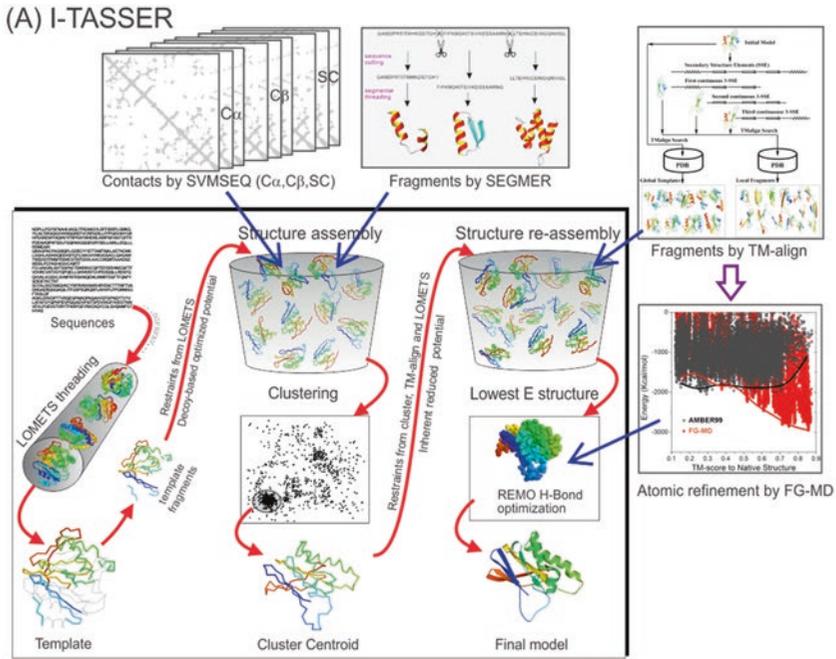
3 Methods

3.1 Inferring Structure and Function of Protein Isoforms Using Structural Bioinformatics Tools

We developed an integrated computational approach with tools for 3D protein structure prediction and structure-based protein function annotation, which can be used for inferring potential folding and functions of the noncanonical splice proteins compared with those of the canonical proteins. As examples, we first applied this approach to examining the function and folding stability of pyruvate kinase M2 isoforms, whose 3D structures were known, and three cancer-related isoform pairs, Bcl-x, caspase 3, and odd-skipped related 2, which were reported to have opposite functions, but lacked experimentally derived structures [5].

3.1.1 Protein 3D Structure Prediction

We developed and recommend two methods, I-TASSER [6–11] and QUARK [12–14] for template-based and ab initio protein structure predictions, respectively (Fig. 1).



(B) QUARK

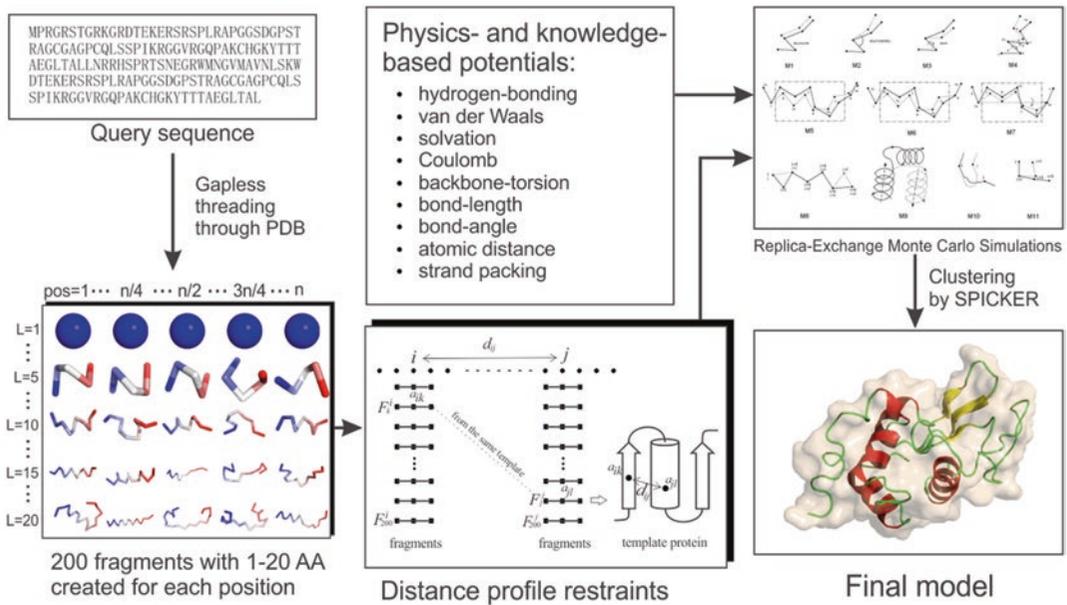


Fig. 1 Flowcharts of template-based structure prediction by I-TASSER (a) and ab initio structural folding by QUARK (b)

For additional comments about the I-TASSER family of tools and alternative methods, *see* Subheading 4.

I-TASSER pipeline for template-based structure prediction. The pipeline of I-TASSER is presented in Fig. 1a. The classic I-TASSER is depicted in the left-bottom panel of Fig. 1a, which consists of three steps:

1. Thread the query sequence through the PDB library by a locally installed meta-threading-server (LOMETS) method [15] to identify structure templates. Because LOMETS combines multiple threading programs based on different but complementary alignment algorithms, the templates often have improved structural coverage and alignment accuracy compared to individual threading algorithms.
2. Construct full-length model by reassembling the continuous segments excised from the templates in the threading-aligned regions using iterative replica-exchange Monte Carlo (REMC) simulations, where the structure in the threading-unaligned regions is built by an on-lattice ab initio modeling procedure. The REMC simulations are guided by a highly optimized knowledge-based force field that consists of generic backbone contact and correlation interactions, hydrogen bonding, and threading template-based restraints [11, 16].
3. Select the lowest free energy models by the clustering of all structure decoys as generated by the REMC simulations [17, 18]; the atomic models are then refined by REMO via the optimization of the hydrogen-bonding networks [19].

In recent developments, we proposed three methods to improve the accuracy of I-TASSER for distant homology modeling. These new developments are depicted in the upper-right panel of Fig. 1a, which was proved efficient in improving the modeling accuracy for proteins that lack close homologous templates [20]. These include:

1. Extended SVMSEQ [21] to generate multiple-scale residue contacts to guide the long-range I-TASSER structure assembly [22]
2. Developed SEGMER [23] to detect super-secondary structural motifs by segmental threading which improve spatial restraints of medium-range I-TASSER simulations
3. Applied FG-MD [24] to refine I-TASSER models at the atomic level based on fragment-guided molecular dynamic simulations, where the fragments from the PDB are shown to be able to improve the funnel of landscape of the physics-based force field.

QUARK pipeline for ab initio protein structure folding. Because the accuracy of I-TASSER predictions often relies on the existence of PDB templates and cannot be used to model proteins with new folds, we developed QUARK for ab initio protein structure prediction (Fig. 1b), [12, 25], which consists of three steps:

1. Select 200 short fragments at each position of the target sequence, each with 1–20 residues, from unrelated proteins by gapless matches.
2. Use the selected fragments to assemble full-length models by REMC simulations, under a composite physics- and knowledge-based potential that consists of H-bond, van der Waals, solvation, Coulomb, backbone torsion, bond length and angle, and statistical strand and helix packing interactions [12]. Meanwhile, non-covalent contact and distance profiles are derived from the short-range fragments by consistency analysis and used to accommodate long-range packing simulations [25].
3. Select final models by SPICKER clustering program, and then refine the atomic models by ModRefiner [26] through a two-step atomic-level minimization to improve H-bond networks and physical realism.

Test of structure prediction methods in blind CASP experiments. I-TASSER and QUARK have been tested in both benchmark [6, 7, 12, 13, 27] and blind experiments [8, 9, 10, 28–30]. For the blind test, community-wide CASP experiments have been organized every two years since 1994 to examine the state of the art of structure prediction methods (<http://predictioncenter.org>) [31, 32]. Structure predictions of a set of >100 protein targets are made *before* the experimental structures are released, and the modeling results by predictors are assessed by independent scientists. The CASP experiments have attracted hundreds of predictors from the community in the last two decades. I-TASSER was tested (as “Zhang-server”) in the seventh–eleventh CASP competitions in 2006–2014, and QUARK participated in the ninth and tenth CASPs. I-TASSER and QUARK have consistently ranked in the top two positions in the automated server section for generating the most accurate protein structure predictions [29, 30, 33–35]. These results demonstrate the advantage of these pipelines over other state-of-the-art methods for high-resolution protein structure predictions.

3.1.2 Structure-Based Function Annotations

To annotate the biological functions of proteins, we first developed a high-quality protein function database, BioLiP [36], semi-manually curated from databases and PubMed literature. Two complementary approaches, COFACTOR [37, 38] and COACH [39], were then proposed to predict the protein functions by structurally matching the prediction structure models with the known proteins in BioLiP.

Development of protein functional databases. Many structure-based protein function analyses and prediction studies use known proteins solved in the PDB [40] as templates to infer biological functions of unknown proteins. However, numerous proteins in the PDB contain redundant entries and/or misordered residues

and functions. In particular, many proteins were solved using artificial molecules as additives to facilitate the structural determination experiments. These ligands do not necessarily represent biologically relevant binding. Therefore, it is essential to develop cleaned protein libraries with biological functions carefully validated. We proposed a hierarchical procedure, which consists of three steps of computational filtering and manual literature validation for assessing the biological relevance of the annotated protein functions.

1. Download 3D structure for each entry in the PDB, with the modified residues (i.e., residues modified post-translationally, enzymatically, or by design) translated to their precursor standard residues based on the MODRES record. To exclude crystallization neighbors, the biological unit files rather than the asymmetric unit files are used for evaluating the ligand-protein contacts.
2. Extract ligands, which are defined as small molecules, from the PDB file. Three types of ligand molecules are collected in the BioLiP database, the molecules from the HETATM records (excluding water and modified residues), small DNA/RNA, and peptides with less than 30 residues. If the closest interatomic distance between two HET group ligands is smaller than 2 Å, the two ligands are merged as a single ligand and are regarded as a *k*-mer ligand.
3. Submit each ligand molecule to a composite automated and manual procedure to decide its biological relevance. If the ligand molecule is evaluated as biologically relevant, its interaction with the receptor (i.e., binding site residues in the receptor) is deposited into the BioLiP database. Additionally, the ligand-binding affinity, catalytic site residues, EC numbers, GO terms, and cross-links to the PDB, UniProt, PDBsum, PDBe, and PubMed databases are also collected and deposited into BioLiP.

BioLiP is updated weekly and is freely available for the community at <http://zhanglab.ccmb.med.umich.edu/BioLiP>. The current release of BioLiP contains 344,990 entries constructed from 72,005 unique PDB proteins, in which 40,078 entries are for DNA/RNA-protein interactions, 15,648 for peptide-protein interactions, 94,907 for metal ion-protein interactions, and 184,357 for regular small molecule-protein interactions. There are in total 23,492 entries with binding affinity data collected from Binding MOAD [41], PDBbind-CN [42], and BindingDB [43] databases and from a manual survey of the literature. It also contains proteins of known enzyme commission (EC) [44] and gene ontology (GO) [45]. Currently, the EC domain of BioLiP involves 7674 protein chains with 203 unique first three-digit and 1900 unique four-digit enzyme commission numbers. The GO domain contains 26,004 chains/domains associated with 11,686 unique

gene ontology terms. These data provide important resources for function annotation studies.

COFACTOR method for EC, GO, and ligand-binding predictions. COFACTOR is a template-based approach developed for structure-based function annotation [37, 38].

1. Identify functional templates by global and local structural matches of the target proteins with the known proteins in the BioLiP [36]. The target structures can be either from experimental determination or computational structure prediction.
2. Infer functional insights, including ligand-binding sites, enzyme commission number, and gene ontology terms, from the highest scoring function templates.

COFACTOR was tested in the function prediction section of the ninth CASP experiment, where COFACTOR (as “I-TASSER_FUNCTION” in the server section and “ZHANG” in the human section) was ranked in the first two positions [46].

COACH protocol for ligand-binding prediction. Because individual methods can only generate predictions for specific protein targets, we proposed a new protocol, COACH, which aims to extend the coverage of function predictions by combining results from multiple methods (Fig. 2) [39]. The pipeline consists of two steps:

1. Use two complementary algorithms TM-SITE and S-SITE to derive functional templates from BioLiP library. TM-SITE was designed to recognize functional templates by binding-specific substructure comparison combined with biochemical feature alignment of binding pockets. S-SITE uses protein sequence profile collected by PSI-BLAST that is then searched through BioLiP.

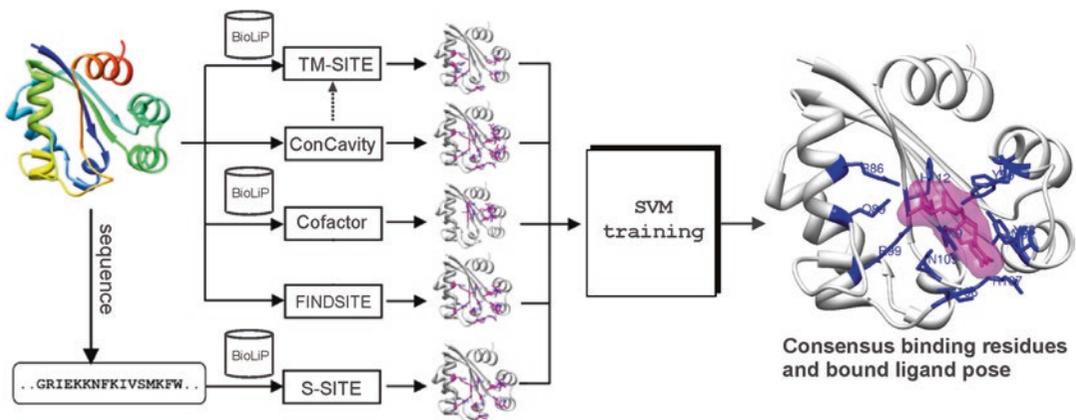


Fig. 2 Protocol of COACH for structure-based protein function annotation. Prediction results from five different programs are combined using support vector machine to increase accuracy and coverage of function annotations

2. Combine multiple predictions by TM-SITE and S-SITE, together with three other predictors (COFACTOR [38], FINDSITE [47], and Concavity [48]), to derive consensus function annotations using support vector machine (Fig. 2).

The COACH protocol was examined in the recent community-wide COMEO experiment designed to test ligand-binding predictions using prereleased PDB sequences as a continuous base [49]. COACH was consistently ranked as the best method in the last 22 individual datasets with its average AUC score, the area under the curve of the true positive rate versus false positive rate plot, 22.5 % higher than the second best method [49].

3.1.3 Structure-Based Function Annotations of Alternative Splice Proteins Expressed in Her2/Neu-Induced Breast Cancers

Identification of splice variant peptides in tumor tissue of mice with her2/neu-amplified breast cancer. We analyzed tumor and normal mammary tissue LC-MS/MS datasets from the Chodosh mouse model of Her2/neu-driven breast cancer, which accounts for 15–20 % of breast cancers in humans [50]. A total of 608 distinct alternative splice variants, 540 known and 68 novel, were identified [3]. There were 216 more from the tumor lysate than from the normal sample (505 vs. 289), probably reflecting greater cellularity and higher expression per cell. We chose 32 of the 45 novel proteins expressed only in tumor specimens for confirmation with qRT-PCR; all were confirmed except for one primer which did not work, and 29 of 31 showed increased mRNA expression. Of the 15 biomarker candidates that Whiteaker et al. [50] confirmed as over-expressed in tumor lysates with MRM-MS, we found that 10 had splice variants in our analysis, although we had no information on the functional activities of the different isoforms of these or any other proteins from proteomic analyses.

Among the 68 novel proteins, we demonstrated variants resulting from new translation start sites, new splice sites, extension or shortening of exons, deletion or swap of exons, retention of introns, and translation in an alternative reading frame. Our annotations revealed multiple variants with potential significant functional motifs, including two relating to BRCA1 through binding to its BRCT domain. The peptide sequence “FSRAEAE GPGQACPPRPFC” is in the second intronic region of leucine zipper-containing LF (Rogdi) gene. Using Splice Site Prediction by Neural Network from the Berkeley Drosophila Genome Project (http://w.fruitfly.org/seq_tools/splice.html), we found a predicted donor splice site “gactgaggtgaggtg” where the novel peptide was identified as coding sequence with a splice site prediction score of 0.93. Functional motifs identified in expressed intronic sequences include LIG_BRCT_BRCA1_1, a phosphopeptide motif which interacts directly with the carboxy-terminal domain of BRCA1. The peptide “GSGLVPTLGRGAETPVSGAGATRGLSR” aligned to the first intronic region of transcription factor *sox7*; the very same LIG-BRCT_BRCA1_1 motif was found in this intronic region.

Structure modeling-based analysis on alternative splice proteins in Her2/neu-induced breast cancers. Experimentally determined structures of protein splice isoforms remain rare; in fact, there were only seven full-length pairs of such isoforms in the enormous Protein Data Bank (PDB) and the Alternative Splicing and Transcript Diversity (ASTD) database as of 2011 [5]. Homology modeling methods are poor at predicting atomic-level structural differences because of the high sequence identity between the isoforms. We exploited the bioinformatics tools [6, 11, 38, 39] described above (Subheading 3.1) to analyze the folding, structural conformation, and likely functional consequences of alternative splicing of proteins identified in the Her2/neu-induced breast cancer model [5]. The procedure of the application consists of three steps:

1. Based on the I-TASSER modeling, we have demonstrated that attributes in ab initio structural assembly and template refinement can partially differentiate atomic details of splice protein variant pairs [5]. The structure modeling approach was benchmarked on the seven pairs of protein splice isoforms with solved structures in PDB, which resulted in structural models with an average RMSD = 1.72 Å to the native, after excluding all homologous templates to the targets. Some of the structural variations in the isoform pairs were due to exon swapping. Even alternative splice variants whose structures are very similar may have functional differences due to the absence of a functionally critical residue or altered post-translational modifications of residues in the swapped exon. For example, in the case of acid phosphatase (acp1) variants, the Mg²⁺-binding site is missing in the 1xwwA variant.
2. We used the strategy to model three cancer-related variant pairs reported to have opposite functions, but lacking experimentally derived structures: Bcl-x, caspase 3, and odd-skipped related 2. In each isoform pair, we observed structural differences in regions where the presence or absence of a motif can directly influence the distinctive functions of the variants. For example, an additional 63 amino acids (AA 129–191) create an extra domain in the core structure of bclx-L (233 AA) compared with bclx-S (170 AA); the shorter variant is missing the two Bcl-2 family motifs BH1 and BH2, while the longer variant contains all four Bcl-2 homology motifs (BH1–4). This difference results in a completely different topology and function; bclx-L is antiapoptotic, while bclx-S is proapoptotic.
3. We applied I-TASSER to five splice-variant pairs overexpressed in the mouse Her2/neu mammary tumor: annexin 6, calumenin, cell division cycle 42 (cdc42), polypyrimidine tract-binding protein 1 (ptbp1), and tax1-binding protein 3 (tax1bp3). These pairs were chosen based on the following five criteria: differential expression, annotated as a known protein in Ensembl,

at least 75 % sequence identity with the canonical protein, known homologous variants of the protein pair in *Homo sapiens*, and an I-TASSER confidence score (C-score) for both variants >-1.5 to ensure the quality of structure prediction.

Despite the high sequence identity between the variant pairs (99, 92, 95, 95, and 79 %, respectively), structural differences were revealed in biologically important regions of these protein pairs. For example, the only difference between anxa6-001 and anxa6-002 at the sequence level is the presence of six residues in anxa6-001 (VAAEIL, AA 525–530) that are missing in anxa6-002. The global topology of the I-TASSER models of the two isoforms is almost identical, with RMSD = 0.38 Å and TM-score = 0.99. However, there is an obvious local structural change in the region due to the absence of “VAAEIL” residues (AA 525–530 in anxa6-001), as identified by the structural alignment algorithm, TM-align [51]. As reported, these six residues are in the end of a helical region (blue-colored in the original figure) which is followed by a loop. Because of the absence of the six residues, the loop is smaller in the shorter variant. The nearby proline-directed kinase phosphorylation ([ST]P) site followed by a serine phosphorylation site moves from 535–537 to 529–531, inside the helix region in anxa6-002, where phosphorylation is less probable than for anxa6-001.

The I-TASSER models in Fig. 3 show that the threonine and proline residues are buried by other atoms in the anxa6-002

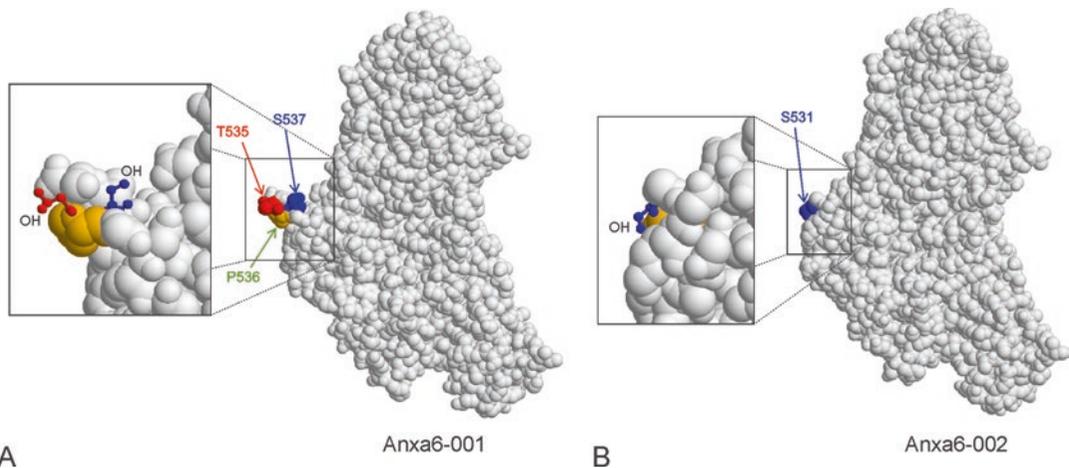


Fig. 3 The I-TASSER models for two splice isoforms of Annexin 6. (a) The “TPS” residues in anxa6-001 are exposed to solvent which helps increase the likelihood of phosphorylation as a post-translational modification. The hydroxyl groups, which are the target of kinases for phosphorylation, are highlighted in the inset. (b) Due to the absence of “VAAEIL” residues (aa 525–530 in anxa6-001) in the anxa6-002 variant, the “TPS” residues are either partially or completely buried by other atoms which significantly reduce the possibility of the protein for phosphorylation [Modified from reference [1]]

variant, whereas the hydroxyl group of serine (S531 in anxa6-002) seems quite accessible for phosphorylation (see inset of Fig. 3b). In order to search for phosphopeptides from the spliced region of anxa6, we performed a fresh analysis of the mass spectrometric data with our custom database using X! Tandem, specifying phosphorylation on serine or threonine (phospho(S) and phospho(T)) as potential residue modifications. Because phosphorylation is usually present at low stoichiometry and our dataset was not enriched for phosphopeptides, it was striking that we identified a spectrum from the normal sample that matched to the peptide “DQAQED A QVAAEILEIADTPSGDKTSLETR” with 3281.5 daltons as the calculated peptide mass plus a proton (mh). The unmodified mh of this peptide is 3201.5 daltons; the additional 80 daltons can be accounted for precisely by phosphorylation of either the threonine or serine residue in the peptide. We did not find such a phosphopeptide from the tumor sample. However, we did find multiple high-quality spectra from the tumor sample that identified the sequence “DQAQEDAQEIADTPSGDKTSLETR,” the unique peptide that matches the anxa6 short variant (with residues “VAAEIL” missing). None of these spectra revealed modification by phosphorylation. Even though only a single spectrum identified the phosphopeptide from the unique region of the long anxa6 variant in the X! Tandem search, these observations are consistent with our functional inference from the structural comparison of the anxa6 variants [5] that the longer anxa-001 variant is more prone to undergo phosphorylation at Thr-535 or Ser-537 than in the anxa-002 variant at the Thr-529 or Ser-531 sites. Post-translational phosphorylation of anxa6 has been reported to be associated with cell growth in 3T3 fibroblasts and human T-lymphoblasts [52]; we previously predicted that the critical phosphorylation may occur at Thr-535 and/or Ser-537 in the loop region. We have now strikingly refined this prediction, which we hope experimentalists will test.

3.2 Methods for Annotation of Protein Isoform Structure and Predicted Functions, Using Support Vector Machine Multiple-Instance Learning to Predict Functions of Isoforms and Isoform-Level Networks

The prediction of isoform functions and networks can be formulated as a problem that can be addressed by multiple-instance learning (MIL) algorithms [53]. MIL works mainly in three steps (Fig. 4):

1. Initialization: label all isoforms in positive genes as Class 1 and all isoforms in negative genes as 0.
2. Model building: build a classifier using a given method such as support vector machines and Bayesian networks followed by using this classifier to score all isoforms.
3. Witness updating: reselect the highest scored isoform from positive genes as “witness,” and label them as Class 1. All other isoforms are labeled as Class 0. If results are not converged, go to **step 2**. Otherwise, the converged model is stored for predicting isoform functions.

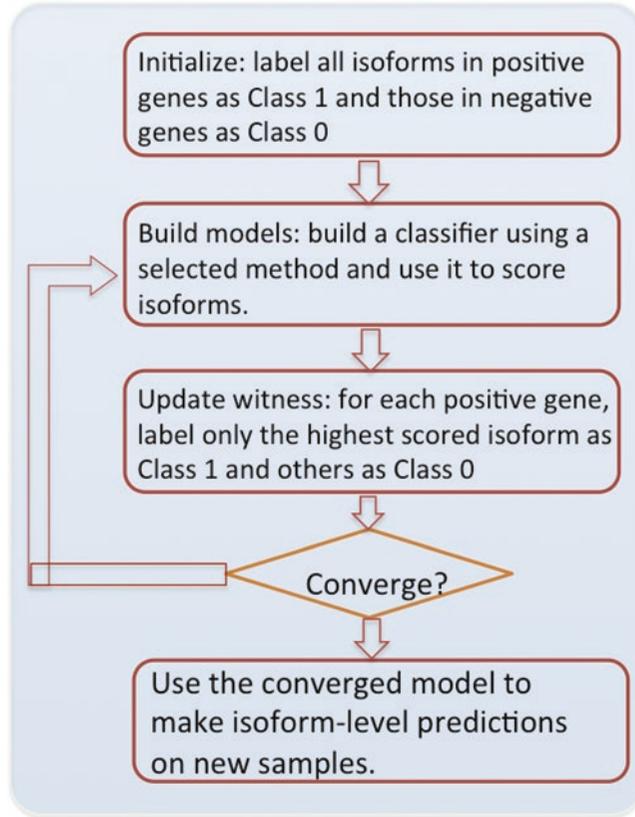


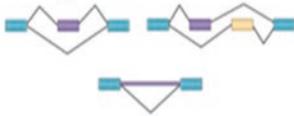
Fig. 4 The schematic of iterative multiple-instance learning (MIL) for isoform-level function prediction

3.2.1 Genome-Wide Isoform Functions and Networks in the Mouse

Eksi et al. performed the first genome-wide isoform function predictions using the MIL algorithm for the mouse [53]. The method is described in the following steps:

1. Collect isoform feature data. A total of 811 RNA-Seq experiments (samples) were downloaded from the NCBI Sequence Read Archive (SRA) database (<http://www.ncbi.nlm.nih.gov/sra>), followed by estimating isoform expression using the TopHat and Cufflinks suites (v2.0.051). Quality control was conducted [53], and 365 samples were kept for subsequent analysis. All isoform expression from these 365 samples was collected into a data matrix.
2. Select “gold standard” of gene functions. Biological process terms in GO were used as “gold standard” for functional annotation.
3. Learn a support vector machine (SVM) model from the above collected data using the MIL algorithm (Fig. 4).
4. Use the learned SVM model to predict functions for all mouse isoforms.
5. Build a web server to make isoform functions publicly available and searchable (<http://guanlab.ccmb.med.umich.edu/isoPred>).

Step 1:



ANXA6 SEARCH



Step 2:

Gene Search Results:

Show 25 entries

Gene Symbol

Anxa6

Showing 1 to 1 of 1 entries



Step 3:

You searched for the gene *Anxa6*

GO term	GO term name	Click for neighbors	Sort with maximum	Isoform Name	Fold Change	Isoform Name	Fold Change
GO:0045058	T cell selection	🔗		NM_001110211.1	20.55	NM_013472.4	39.34
GO:0001912	positive regulation of leukocyte mediated cytotoxicity	🔗		NM_001110211.1	7.60	NM_013472.4	32.63
GO:0051650	establishment of vesicle localization	🔗		NM_001110211.1	3.59	NM_013472.4	26.86
GO:0001914	regulation of T cell mediated cytotoxicity	🔗		NM_001110211.1	8.98	NM_013472.4	26.59
GO:0055010	ventricular cardiac muscle tissue morphogenesis	🔗		NM_001110211.1	25.96	NM_013472.4	2.36
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	🔗		NM_001110211.1	2.86	NM_013472.4	24.58

Fig. 5 An example query of *Anxa6* gene from the mouse isoform function database

Taking *Anxa6* (Annexin A6) as an example, the steps for querying the functions of its isoforms are described below (also shown in Fig. 5).

1. Go to website <http://guanlab.ccmb.med.umich.edu/isoPred>, type in “Anxa6” in the input box, and click the “Search” button (step 1, Fig. 5).
2. Users will be guided to http://guanlab.ccmb.med.umich.edu/isoPred/search_result.php. Locate “Anxa6” in the form on this page, and click it (step 2, Fig. 5).

3. Users will be guided to http://guanlab.ccmb.med.umich.edu/isoPred/gene_result.php?gene=Anxa6. On this page, see function predictions for each isoform presented in the table (step 3, Fig. 5).

Further, Li et al. constructed a genome-wide functional relationship network for the mouse [54, 55] with the following steps:

1. Collect isoform-pair feature data: RNA-Seq, exon array, pseudo-amino acid composition (pseudoAAC), and protein-docking data. For RNA-Seq and exon array, isoform expression was estimated followed by calculating isoform correlation as feature data. Correlation was also calculated between isoforms using their pseudoAAC profile. Protein-docking itself is an isoform-pair feature.
2. Construct “gold standard” of functionally related gene pairs. GO biological processes and KEGG and BioCyc pathways were used to construct a “gold standard” of functionally related gene pairs. Genes annotated to the same biological process or GO term are assumed to have a functional relationship.
3. Learn a model using MIL with Bayesian network as the base learner.
4. Use the learned model to make genome-wide predictions of functional relationship between any two isoforms.
5. Build a web server to allow users to search isoform networks. It is publicly available at <http://guanlab.ccmb.med.umich.edu/isoformnetwork>.

Users can go to this server, input their gene(s) of interest, click the “Search” button, and see isoform networks along with GO enrichment results.

Based on the mouse isoform network, Li et al. catalogued the highest connected isoforms (HCIs) as a predicted “canonical isoform” using the following approach [55].

1. Calculate an average functional relationship (AFR) score for each isoform of multi-isoform genes.
2. For each multi-isoform gene, choose the isoform with the highest AFR score as HCI. The remaining isoforms are considered as NCI (non-highest connected isoforms).
3. Use independent RNA-Seq and proteomic data to investigate the expression of HCI at both transcript and protein levels.
4. Identify a set of genes whose HCIs are most expressed at transcript level and are also expressed at protein level.

Further, the MISOmine database was developed to provide an integrated platform for analyzing isoform expression, functions,

and networks for the mouse [56]. Users can go to the website (<http://guanlab.ccmb.med.umich.edu/misomine/>) to perform isoform-level analyses.

3.2.2 Genome-Wide Isoform Functions and Networks in the Human

Panwar et al. predicted functions for splice isoforms in humans [57]. The approach mainly consists of the following steps:

1. Download the human RNA-Seq data from the ENCODE study [58]; 127 samples were used. The TopHat and Cufflinks suites were used to estimate isoform expression in terms of FPKM [59] using the Ensembl gene annotation (version74, available at <http://www.ensembl.org/>).
2. Use gene ontology biological processes to construct “gold standard” functional annotations.
3. Build an SVM model using the MIL algorithm, and use the model to predict the functions of all human isoforms.
4. Build a web server to store all the predictions and to make the predictions searchable (<http://guanlab.ccmb.med.umich.edu/isofunc/>).

In addition to isoform functions, Li et al. built a genome-wide function relationship network at the isoform level for the human [60], an effort from the chromosome 17 Human Proteome Project [61]. The pipeline is described below:

1. Collect four types of isoform-level feature data, including RNA-Seq expression, pseudo-amino acid composition, protein-docking score, and conserved domains.
2. Construct a “gold standard” of functionally related gene pairs using GO biological processes and KEGG pathways.
3. Build a Bayesian network classifier using the MIL algorithm, and use the model to make genome-wide predictions of function relationships between isoforms.
4. Build a web server to store the human isoform network (<http://guanlab.ccmb.med.umich.edu/hisonet>). Users are able to query isoform networks of their genes of interest.

In addition, using the same method described in Subheading 3.2.1, Li et al. catalogued a set of HCIs for the human.

3.2.3 Comparison of Isoform Functions and Networks in Mice and Humans

Li et al. compared the HCIs between mouse and human to investigate whether they are conserved:

1. Choose a set of 306 multi-isoform homologous genes between mouse and human. For each of these genes, denote its mouse and human HCI as HCI_m and HCI_h .
2. Identify 61 of the 306 genes whose HCI_m and HCI_h are homologs based on the HomoloGene database in NCBI (<http://ncbi.nlm.nih.gov/homologene>).

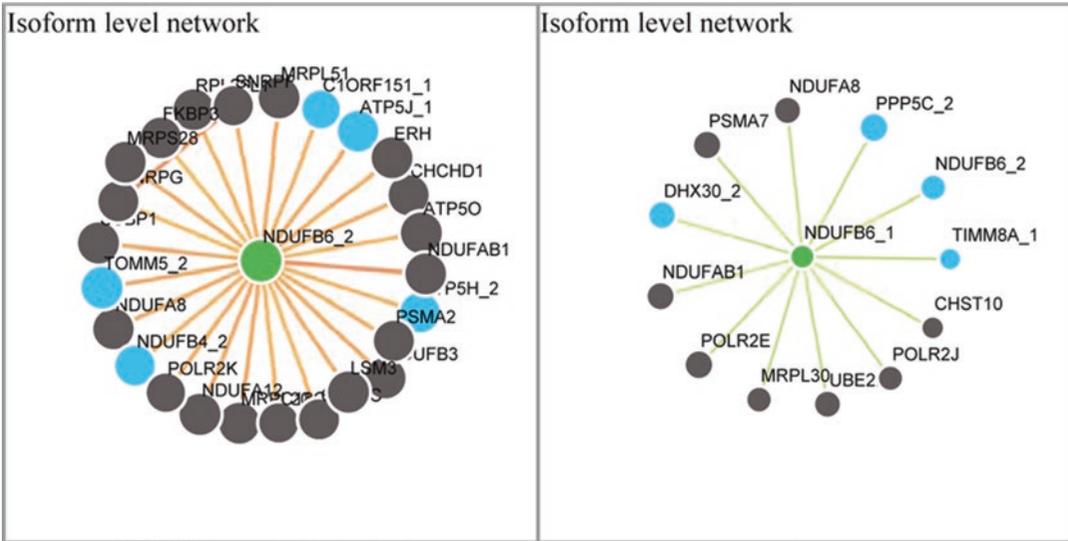


Fig. 6 The functional networks of the highest connected isoform (HCI) (NM_002493.4, NDUFB6_2) and non-highest connected isoform (NCI) (NM_182739.2, NDUFB6_1) of the NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6 (NDUFB6) gene

3. Computationally generate a null distribution of the number of genes whose HCI_m and HCI_h are homologs by chance (41 ± 6).
4. Calculate a p-value by comparing the observed value [61] to the null distribution.

The overlap between mouse and human HCI is significant ($p = 0.0003$), providing additional evidence supporting the “canonical” features of HCIs. As an example, *NDUFB6* is a gene encoding two splice isoforms: HCI (NM_002493.4, NDUFB6_2) and NCI (NM_182739.2, NDUFB6_1); their networks are shown in Fig. 6. The HCI, but not the NCI, was reported to be expressed at the protein level in normal human retina [60], further supporting the “canonical” characteristics of HCIs.

3.2.4 Examples from the ERBB2 Amplicon and Pathways

The ERBB2 (HER2) gene is an epidermal growth factor (EGF) receptor of the receptor tyrosine kinase family. The protein encoded by this gene functions by forming a heterodimer through binding with ERBB1, ERBB3, or ERBB4 receptor proteins. Amplification or overexpression of this gene has been shown to be strongly associated with a major subset of human breast cancers. One can explore the isoform functions and networks for this gene:

1. Isoform functions. From the IsoFunc database (<http://guanlab.ccmb.med.umich.edu/isofunc/>), identify the functions of the five protein-coding splice isoforms of ERBB2 based on Ensembl annotation (version 74). For example, ENST00000541774 was

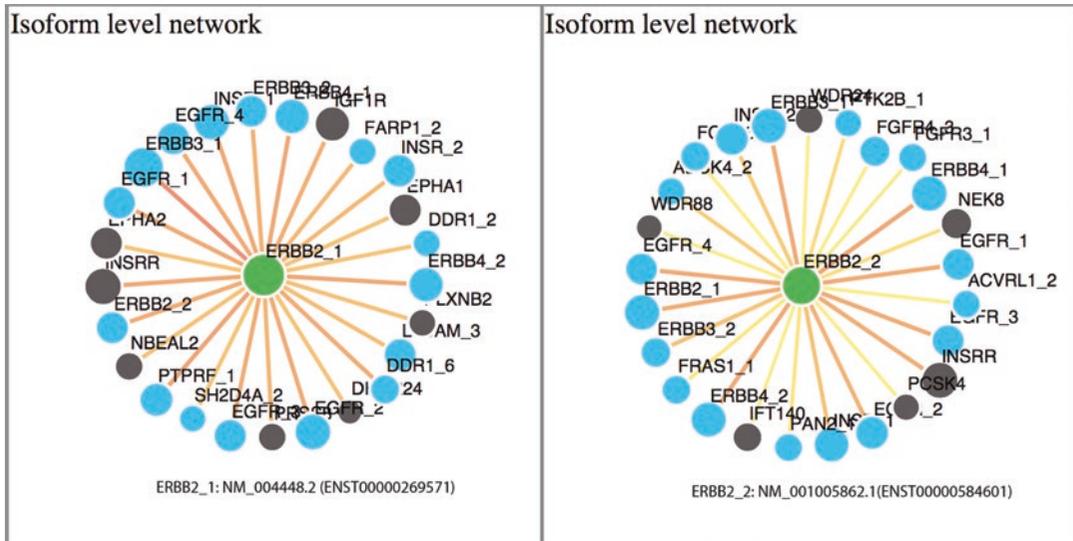


Fig. 7 Isoform networks of ERBB2_1 (NM_004448.2) and ERBB2_2 (NM_001005862) of the human ERBB2 gene

predicted to carry the “canonical Wnt signaling pathway” with a fold change = 37. In contrast, the likelihood for the other isoforms to have this function is much smaller (fold change ranging from 1.7 to 3.3). ENST00000445658, ENST00000269571, ENST00000584601, and ENST00000406381 are predicted to most likely function in “sterol biosynthetic process,” “extracellular matrix disassembly,” “extracellular matrix disassembly,” and “cell-substrate junction assembly,” respectively. These predictions suggest the potential functional differences between isoforms, though they need to be experimentally validated.

2. Isoform networks. To explore further the functional interactions, obtain the isoform networks of ERBB2 isoforms from <http://guanlab.ccmb.med.umich.edu/hisonet/>. This server provides networks for two isoforms, NM_004448.2(ENST00000269571) and NM_001005862.1(ENST00000584601), which are shown in Fig. 7. Both networks are enriched for GO biological processes such as “phosphoinositide 3-kinase cascade” and “protein amino acid autophosphorylation,” suggesting their functional similarities. Functional differences were also predicted. For instance, GO term “regulation of MAP kinase activity” was enriched in the network of NM_004448.2 but not NM_001005862.1, suggesting possible different roles/extent for ERBB2 isoforms to be involved in the MAP kinase signaling pathway.

3.3 Concluding Remark

The combination of proteomics and transcriptomics with the bioinformatics algorithms and methods of structural biology and functional relationship networks can generate many new insights and provide testable hypotheses for experimental studies.

4 Notes

There are a variety of computational tools that have been developed in the field, including, e.g., Rosetta [62], HHsearch [63], and Modeller [64] for protein structure prediction and Concavity [48], FINDSITE [47], and ProFunc [65] for structure-based function annotations. While the I-TASSER family tools represent one of the most efficient sets of methods as demonstrated in various community-wide structure and functional modeling experiments [29, 34, 46, 66], it is important to remember that the results are predictions from automated computational programs. The accuracy and confidence of the models vary among different proteins, depending on the availability of homologous templates and size of the target sequences. We have developed two confidence scores to guide their use by biologist users.

First, C-score [27] is a measurement of confidence of protein structure models built by I-TASSER [11] and QUARK [12] programs. It was defined based on the significance score of structure templates identified by threading alignments and the structural density of Monte Carlo-based conformational search. A large-scale benchmark experiment based on 500 nonredundant proteins showed that there is a high correlation between the C-score and TM-score of the predicted models, with a Pearson correlation coefficient = 0.91 [27].

Second, we proposed an F-score [38] to estimate the accuracy of structure-based function predictions by COFACTOR [37] and COACH [39]. The F-score was defined based on the C-score of protein structure predictions and the structural and sequence similarities between the target and template proteins. A positive correlation between F-score and the accuracy of the predicted models was found in both COFACTOR and COACH predictions.

The I-TASSER family tools have been designed to predict protein structure and functions from the primary sequences. However, information from experimental data or human-based functional analyses can be of critical importance to improve the accuracy of the modeling. The on-line servers and downloadable packages of the I-TASSER family tools have provided entries that allow users to conveniently introduce experimental constraints, including contact and distance maps and specific template alignments, to the modeling systems.

There are multiple factors that would affect the prediction of functions and networks and thus subsequent comparisons, such as choice of gene annotation software for estimating isoform expression. For example, the predictions of human isoform functions [57] and networks [60] are based on RefSeq (version 37.2) and Ensembl (version 74) gene annotations, respectively, so preliminary interpretation of comparative results should be viewed with caution. RefSeq annotation is of high quality but is much less

complete compared to Ensembl, which contains many more (predicted) genes and isoforms. This annotation difference will affect the estimation of splice isoform expression and the subsequent prediction of functions and networks. Also, note that Hisonet provides functions based on GO enrichment, which is different from the directly predicted isoform functions in IsoFunc.

References

1. Omenn GS, Menon R, Zhang Y (2013) Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. *J Proteomics* 90:28–37
2. Menon R, Panwar B, Eksi R, Kleer C, Guan Y, Omenn GS (2015) Computational inferences of the functions of alternative/noncanonical splice isoforms specific to HER2+/ER-/PR-breast cancers, a chromosome 17 C-HPP study. *J Proteome Res* 14(9):3519–3529
3. Menon R, Omenn GS (2010) Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Res* 70(9):3440–3449
4. Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* 32(5):462–464
5. Menon R, Roy A, Mukherjee S, Belkin S, Zhang Y, Omenn GS (2011) Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers. *J Proteome Res* 10(12):5503–5511
6. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4):725–738
7. Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 5:17
8. Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* 69(S8):108–117
9. Zhang Y (2009) I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins* 77(S9):100–113
10. Zhang Y (2014) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* 82(Suppl 2):175–187. doi:[10.1002/prot.24341](https://doi.org/10.1002/prot.24341).
11. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 12(1):7–8
12. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 80(7):1715–1735
13. Xu, D, Zhang, Y (2012) Towards optimal fragment generations for ab initio protein structure assembly. *Proteins*. 10.1002/prot.24179.
14. Xu D, Zhang J, Roy A, Zhang Y (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* 79(Suppl 10):147–160
15. Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucl Acids Res* 35:3375–3382
16. Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J* 85:1145–1164
17. Zhang Y, Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* 25(6):865–871
18. Swendsen RH, Wang JS (1986) Replica Monte Carlo simulation of spin glasses. *Phys Rev Lett* 57(21):2607–2609
19. Li Y, Zhang Y (2009) REMO: a new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins* 76(3):665–676
20. Zhang Y (2014) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* 82(Suppl 2):175–187
21. Wu S, Zhang Y (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* 24(7):924–931
22. Wu S, Szilagy A, Zhang Y (2011) Improving protein structure prediction using multiple sequence-based contact predictions. *Structure* 19(8):1182–1191

23. Wu S, Zhang Y (2010) Recognizing protein substructure similarity using segmental threading. *Structure* 18(7):858–867
24. Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* 19(12):1784–1795
25. Xu D, Zhang Y (2013) Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* 81(2):229–239
26. Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* 101(10):2525–2534
27. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40
28. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. *Proteins* 69(S8):38–56
29. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T (2007) Automated server predictions in CASP7. *Proteins* 69(S8):68–82
30. Cozzetto D, Kryshtafovych A, Fidelis K, Moulton J, Rost B, Tramontano A (2009) Evaluation of template-based models in CASP8 with standard measures. *Proteins* 77(Suppl 9):18–28
31. Moulton J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* 23(3):ii–iv
32. Moulton J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 15(3):285–289
33. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T (2011) Assessment of template based protein structure predictions in CASP9. *Proteins* 79(Suppl 10):37–58
34. Montelione GT (2012) Template based modeling assessment in CASP10. Paper presented at the 10th community wide experiment on the critical assessment of techniques for protein structure prediction, Gaeta, Italy, 9–12 Dec 2012
35. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV (2016) Evaluation of free modeling targets in CASP11 and ROLL. *Proteins* 84(Suppl 1):51–66. doi:10.1002/prot.24973
36. Yang J, Roy A, Zhang Y (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 41(D1):D1096–D1103
37. Roy A, Zhang Y (2012) Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* 20(6):987–997
38. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res* 40(Web Server issue):W471–W477
39. Yang J, Roy A, Zhang Y (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 29(20):2588–2595
40. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlc A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* 39(Database issue):D392–D401
41. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA (2008) Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res* 36(Database issue):D674–D678
42. Cheng T, Li X, Li Y, Liu Z, Wang R (2009) Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model* 49(4):1079–1093
43. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 35(Database issue):D198–D201
44. Barrett AJ (1997) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur J Biochem* 250(1):1–6
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25–29
46. Schmidt T, Haas J, Gallo Cassarino T, Schwede T (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins* 79(Suppl 10):126–136
47. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* 105(1):129–134
48. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* 5(12):e1000585

49. Schwede T (2015) Montly summary of ligand binding prediction results in CAMEO is at <http://www.cameo3d.org/lb>.
50. Whiteaker JR, Zhang H, Zhao L, Wang P, Kelly-Spratt KS, Ivey RG, Piening BD, Feng LC, Kasarda E, Gurley KE, Eng JK, Chodosh LA, Kemp CJ, McIntosh MW, Paulovich AG (2007) Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J Proteome Res* 6(10):3962–3975
51. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33(7):2302–2309
52. Moss SE, Jacob SM, Davies AA, Crumpton MJ (1992) A growth-dependent post-translational modification of annexin VI. *Biochim Biophys Acta* 1160(1):120–126
53. Eksi R, Li H-D, Menon R, Wen Y, Omenn GS, Kretzler MK, Guan Y (2013) Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PLoS Comput Biol* 9(11):e1003314
54. Li H-D, Menon R, Eksi R, Guerler A, Zhang Y, Omenn GS, Guan Y (2013) Modeling the functional relationship network at the splice isoform level through heterogeneous data integration. *bioRxiv*:doi: 10.1101/001719.
55. Li H-D, Menon R, Omenn GS, Guan Y (2014) Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. *Proteomics* 14(23–24):2709–2718
56. Li H-D, Omenn GS, Guan Y (2015) MIsoMine: a genome-scale high-resolution data portal of expression, function and networks at the splice isoform level in the mouse. *Database* 2015. doi: [10.1093/database/bav1045](https://doi.org/10.1093/database/bav1045).
57. Panwar B, Menon R, Eksi R, Li H-D, Omenn GS, Guan Y (2015) Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning under revision
58. Consortium EP (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306:636–640
59. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578
60. Li H-D, Menon R, Govindarajoo B, Panwar B, Zhang Y, Omenn GS, Guan Y (2015) Functional networks of highest-connected splice isoforms: from the Chromosome 17 Human Proteome Project. *J Proteome Res* 14(9):3484–3491
61. Liu SL, Im H, Bairoch A, Cristofanilli M, Chen R, Deutsch EW, Dalton S, Fenyo D, Fanayan S, Gates C, Gaudet P, Hincapie M, Hanash S, Kim H, Jeong SK, Lundberg E, Mias G, Menon R, Mu ZM, Nice E, Paik YK, Uhlen M, Wells L, Wu SL, Yan FF, Zhang F, Zhang Y, Snyder M, Omenn GS, Beavis RC, Hancock WS (2012) A chromosome-centric Human Proteome Project (C-HPP) to characterize the sets of proteins encoded in Chromosome 17. *J Proteome Res* 12(1):45–57
62. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268(1):209–225
63. Soding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951–960
64. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779–815
65. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucl Acids Res* 33(Web Server issue):W89–W93
66. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T (2013) The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database (Oxford)* 2013:bat031

Computational and Statistical Methods for High-Throughput Mass Spectrometry-Based PTM Analysis

Veit Schwämmle and Marc Vaudel

Abstract

Cell signaling and functions heavily rely on post-translational modifications (PTMs) of proteins. Their high-throughput characterization is thus of utmost interest for multiple biological and medical investigations. In combination with efficient enrichment methods, peptide mass spectrometry analysis allows the quantitative comparison of thousands of modified peptides over different conditions. However, the large and complex datasets produced pose multiple data interpretation challenges, ranging from spectral interpretation to statistical and multivariate analyses. Here, we present a typical workflow to interpret such data.

Key words Proteomics, Bioinformatics, Post-translational modifications (PTMs)

1 Introduction

Nearly all proteins have been found to appear in different molecular forms in cellular organisms [1]. Post-translational modifications (PTMs) change the physicochemical properties of proteins affecting their activity, localization, and binding partners. Changes in the PTM abundances and their regulation have a great impact on cellular behavior and aberrant cell states due to diseases [2]. With more than one million human proteoforms estimated to date [3], the characterization of the entire set of proteins comprised in complex samples requires sophisticated experiments, followed by elaborated data analysis [4, 5].

Large-scale PTM analysis experiments based on LC-MS/MS nowadays allow quantifying tens of thousands of modified peptides [6]. While the analysis of intact proteins, top-down proteomics, has great potential to provide unbiased data with complete protein sequence coverage [7], the approach remains analytically challenging for the analysis of complex mixtures [8, 9]. Bottom-up approaches where proteins are proteolytically digested into peptides are thus preferred for high-throughput approaches attempting at identifying and quantifying thousands of modified proteins.

Due to the generally low stoichiometry of modified proteins, their detection relies on various enrichment methods. Different protocols were thus developed allowing the detection of phosphorylation, glycosylation, sumoylation, ubiquitination, acetylation, etc. as reviewed in [6, 10]. As a result, PTM analysis workflows are analytically more complex and costly than other proteomic experiments. Another consequence of this low abundance of modified peptides is that PTM studies require high amounts of starting material, which can become a limitation for rare samples like patient biopsies as reviewed in [11]. Consequently, PTM studies often rely on low number of replicates requiring sophisticated data analysis methods. For a recent review on computational and statistical techniques to tackle data from PTM studies, *see* [12].

Another consequence of the complexity of the workflow often including multiple manual procedures is a high technical variability between samples and high prevalence of missing values. To alleviate these problems, samples are generally labeled and combined, eventually with an internal standard [13–15]. This can be achieved using isobaric tags which allow multiplexing up to ten different samples [16]. Then, isobaric reagents are attached to peptides after digestion, and the samples are pooled in equimolar amounts. The different samples will hence be submitted to the same preparation procedure and remain undistinguishable until mass spectrometry analysis, where the fragmentation of the peptide will produce so-called reporter ions. These ions with masses that are specific to every sample will in turn be used to evaluate the relative abundance of the peptide in the different samples. Two different reagents exist, iTRAQ [17] allowing the multiplexing of four or eight samples and TMT [18] for two, six, or ten samples, as reviewed in [13, 14, 19].

Here, we present a data analysis workflow for such data, where samples for the study of developmental stages of the mouse brain were labeled using iTRAQ 4-plex reagents and enriched for phosphorylation and glycosylation, as well as acetylation. For details, *see* the original publications [20, 21]. We will focus on the analysis of modified and unmodified peptides. Several bioinformatic packages exist for the interpretation of proteomic datasets as reviewed in [13, 14, 19]. However, while standard workflows are well established for global proteome analyses, due to the complexity of the quantitative investigation of PTMs, such studies generally require the application of specialized data interpretation methods [12]. As a result, the interpretation of the datasets generated requires more extensive bioinformatic expertise. We present an example of an interpretation pipeline combining open-source software tools and R scripts. The tools presented are under active development; future versions will thus differ from the ones presented in this Chapter.

2 Materials

2.1 Resources

1. The raw data used in this Chapter as illustrative example are publicly available from the PRIDE database [22] through the ProteomeXchange consortium [23] using the accession (PXD003932). The analyzed samples were extracted from mouse brain tissue at different ages. The samples were treated employing two enrichment strategies yielding three different datasets for each of the four biological replicates (peptides enriched for acetylation, glycosylation + phosphorylation, and without enrichment). For more details on the experimental design, please refer to the original publications [20, 21].
2. The UniProt *Mus musculus* reference proteome database version of the 2 January 2016 (*see Note 1*) is used in this Chapter. In order to reduce the prevalence of false positives (*see Note 2*), contaminant sequences from common Repository of Adventitious Proteins (cRAP, www.thegpm.org/crap) and the sequence of porcine trypsin were appended.

2.2 Software

All software tools used here are freely available and can be downloaded from their respective websites. A detailed tutorial on how to operate these tools can be found at compomics.com/bioinformatics-for-proteomics [24]:

1. Java (java.com): Some of the following software require Java installed (*see Note 3*).
2. ProteoWizard [25] (proteowizard.sourceforge.net): msconvert as part of the ProteoWizard package allows the conversion of raw mass spectrometry files to open formats (*see Note 4*). It can also be used to process the mass spectra [26].
3. SearchGUI [27] (github.com/compomics/searchgui): SearchGUI allows searching spectra using multiple search engines via a user-friendly interface or in command line. At time of writing, SearchGUI supports X! Tandem [28], MyriMatch [29], MS Amanda [30], MS-GF+ [31], OMSSA [32], Comet [33], Tide [34], and Andromeda [35].
4. PeptideShaker [36] (github.com/compomics/peptide-shaker): PeptideShaker can be used to combine and interpret the peptides inferred from spectra by multiple search engines. It is designed to work seamlessly in combination with SearchGUI, via a user-friendly interface or in command line.
5. Reporter (github.com/compomics/reporter). Reporter performs reporter ion-based quantification of data identified by PeptideShaker (*see Note 5*).
6. R software environment (www.r-project.org). R provides a versatile scripting language for statistical data analysis. We furthermore recommend installing RStudio (www.rstudio.com) for improved editing and running of the scripts.

7. R scripts (bitbucket.org/veitveit/ptmanalysis/src). The scripts allow running the quantitative and statistical analysis [37] after simple editing of relevant parameters. Adaptation to different project designs can be achieved by modification of the respective code segments.

3 Methods

3.1 Database Search

1. After downloading and unzipping SearchGUI, double-click on the *.jar* file to start the tool. The mainframe displayed in Fig. 1 allows setting up the search.
2. Select the spectrum files by clicking *Add* next to *Spectrum File(s)*. It is possible to select the raw files directly. *msconvert* will then be automatically enabled (see **Notes 6** and **7**).
3. Select the search settings to use for the search in the drop-down menu next to *Search Settings* (see **Note 8**).
4. In this Chapter, we will create new settings; click on the *Add* button next to the drop-down menu. The dialog displayed in Fig. 2 allows the creation and edition of search settings. Set a name for the search settings in the text field next to *Name* (see **Note 9**). Click on *Show Advanced Settings* to display the options shown in Fig. 2 allowing the setting of advanced identification parameters (see **Note 10**).
5. Click on *Spectrum Matching*; the dialog displayed in Fig. 3 allows to set the search parameters. Select the file containing the protein sequences using the *Edit* button next to *Database (FASTA)*. Select fixed and variable modifications using the arrows in the *Modifications* panel. For this Chapter, use *iTRAQ 4-plex of K* and *Carbamidomethylation of C* as fixed and *Deamidation of N*, *Deamidation of Q*, *Oxidation of M*, and *iTRAQ 4-plex of peptide N-term* as variable (see **Notes 11** and **12**). For fractions enriched for phosphorylation, add *Phosphorylation of S*, *Phosphorylation of T*, and *Phosphorylation of Y*. Set the *Fragment Mass Tolerance* to 0.05 Da in the *Protease and Fragmentation* panel, and let the other settings to default (see **Note 13**). Click on *OK* to finish the edition of the search settings and *Save* to save the identification settings.
6. Select an output folder for the search engine results (see **Note 14**).
7. Select the search engines to use for the analysis; in this Chapter, we use X! Tandem, MyriMatch, and MS Amanda. Advanced search engine parameters can be edited by clicking on the respective cogwheels (see **Note 15**).
8. Go to the *Edit → Advanced Settings* menu and make sure that *Group Identification Files* is set to *Zip File per Mgf* in the *Output* panel.

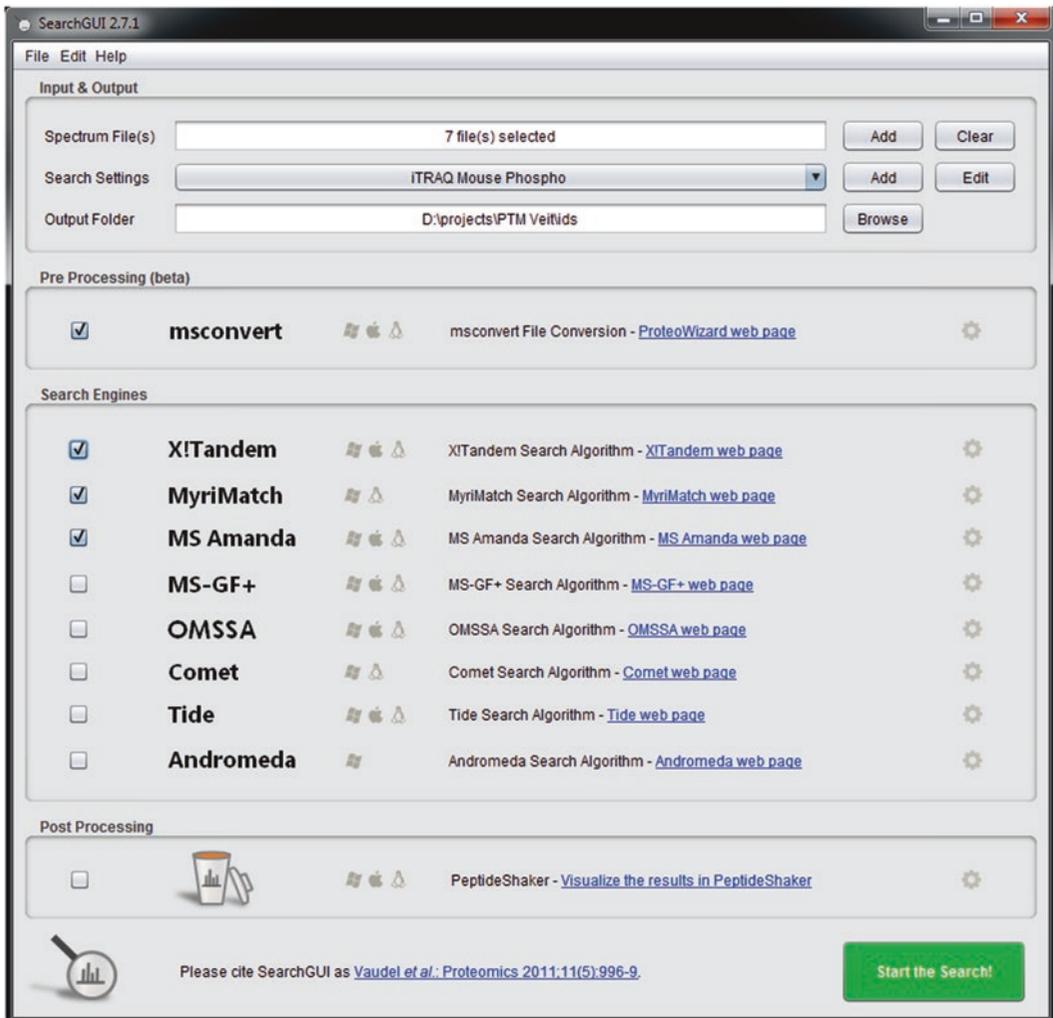


Fig. 1 Main interface of SearchGUI. In the *Input and Output* panel at the *top*, it is possible to provide spectrum files in various formats, select and edit search settings, and set an output folder. *Below*, in the *Pre Processing* panel, msconvert can be operated to convert and process the spectrum files. In the *Search Engines* panel, the desired search engines can be selected. In the *Post Processing* panel, PeptideShaker can be used to combine and interpret the results of the selected search engines. Finally, at the *bottom*, clicking *Start the Search!* launches the process. Note that the settings of every algorithm are available via the cogwheel next to them

9. It is possible to set up the PeptideShaker post processing from the *Post Processing* panel. In this Chapter, the project will be created separately in the next Subheading. Click on *Start the Search!* to launch the process. A dialog appears showing the progress and output of the search engines.

3.2 PSM Aggregation and PTM Localization

1. After downloading and unzipping PeptideShaker, double-click on the *.jar* file to start the tool. In the *Welcome* dialog displayed in Fig. 4, click on *New Project* (see **Note 16**).

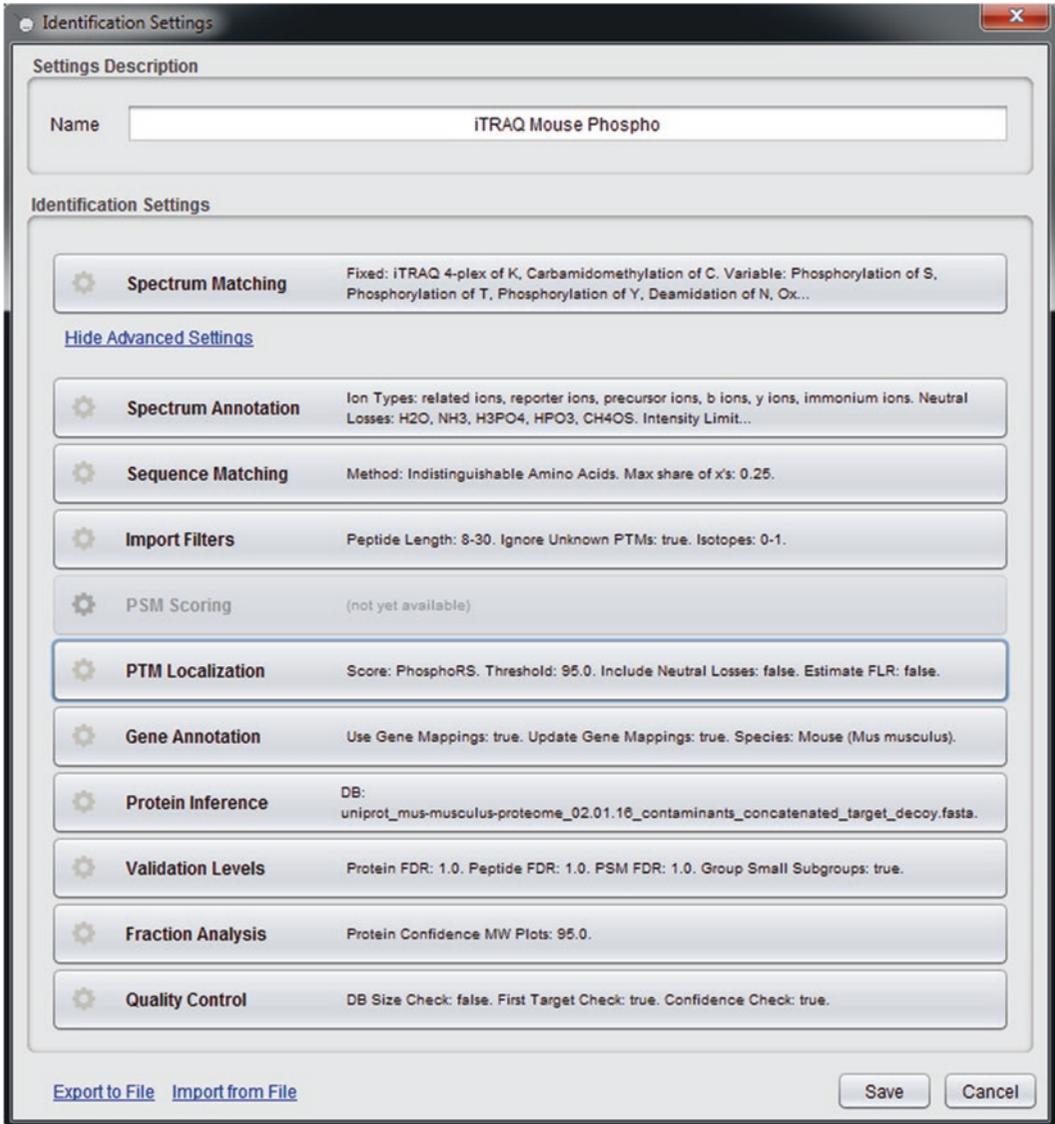


Fig. 2 Identification Settings dialog. In the *Settings Description* panel at the top, it is possible to set a name for the settings. Below, in the *Identification Settings* panel, it is possible to edit the different identification settings. Note that advanced settings are accessible by clicking *Show Advanced Settings*

2. In the project creation dialog displayed in Fig. 5, reference your project in the *Project Details* panel, and select the SearchGUI output corresponding to the files of a given PTM enrichment for a given replicate, e.g., unenriched measurements of replicate two or acetylation of replicate three. SearchGUI output files can be selected via the *Browse* button next to *Identification File(s)* in the *Input Files* panel. Note that the other fields are filled automatically using the output of SearchGUI. If it is not the case, fill these manually.

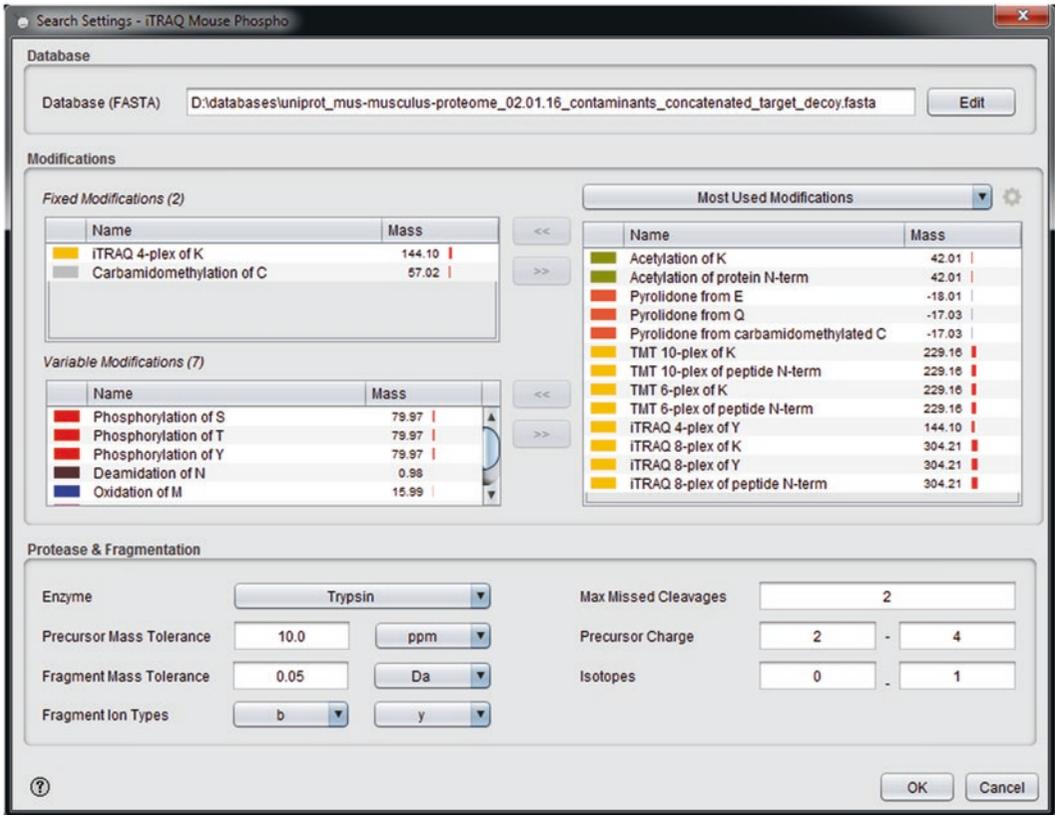


Fig. 3 Search setting dialog. In the *Database* panel at the *top*, it is possible to select the protein sequence database to use for the search. *Below*, in the *Modifications* panel, fixed and variable modifications can be selected. Note that new modifications can be selected using the cogwheel at the *top right*. In the *Protease and Fragmentation* panel, it is possible to set protease used to digest the sample, a given number of allowed missed cleavages, precursor and fragment ion mass tolerances, precursor charge and isotope tolerances, and the type of fragment ions used for the search.

3. Click on *Load Data!*; a dialog will appear showing the progress of the import process.
4. Upon creation, the project can be visualized in the PeptideShaker interface where PTMs are color coded, as displayed in Fig. 6 with the *Overview* tab.
5. Go to the *Modifications* tab. There, it is possible to navigate the peptides carrying the modifications of interest as displayed in Fig. 7.
6. Save the results using the *File* → *Save As...* menu (see **Note 17**).
7. Create a project for every set of files corresponding to a PTM enrichment approach and replicate.

3.3 Reporter Ion Quantification

1. After downloading and unzipping Reporter, double-click on the *.jar* file to start the tool.

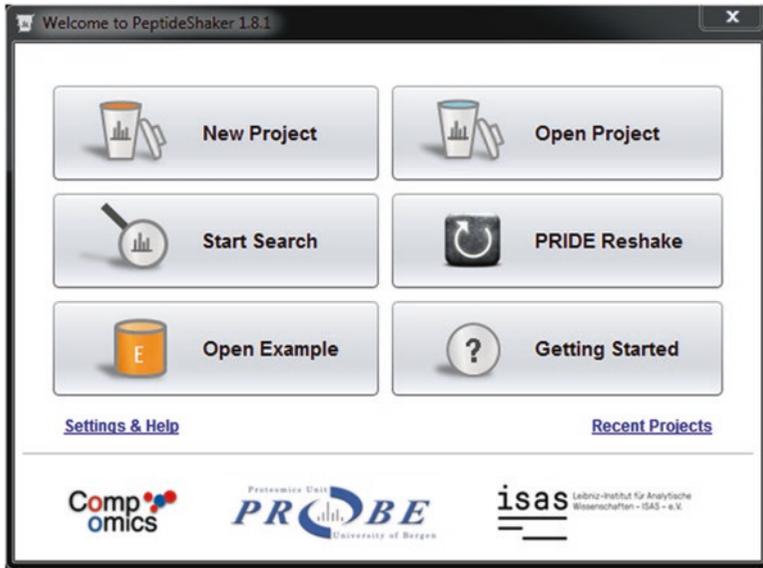


Fig. 4 PeptideShaker Welcome dialog. When starting PeptideShaker, the *Welcome* dialog allows starting a new project, opening a saved project, starting a search using SearchGUI, reprocessing public data from the PRIDE repository, and opening the example dataset or an introductory presentation. At the *bottom*, clicking *Settings and Help* allows setting up software parameters like available memory or temporary folders and provides additional help and the possibility to report bugs



Fig. 5 PeptideShaker project creation dialog. In the *Project Details* panel at the *top*, it is possible to document the project for later reuse. *Below*, in the *Input Files* panel, it is possible to select the search engine result files, the spectrum files, and the database to use for peptide to protein matching. In the *Project Settings* panel, it is possible to select and edit the identification parameters and project and process related settings. At the *bottom*, clicking *Load Data!* starts the data import

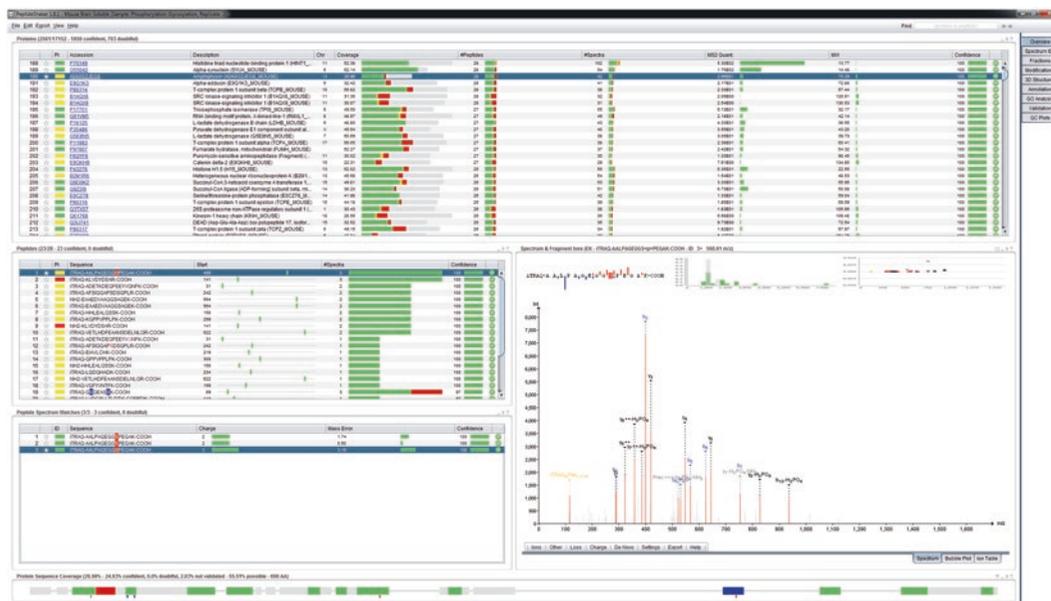


Fig. 6 PeptideShaker Overview tab. The *Overview* tab of PeptideShaker presents the proteomics results in a top-down manner, with proteins listed in the *top table*. The tables to the *left*, respectively, list the peptides of the selected proteins and peptide-spectrum matches (PSMs) of the selected peptide. To the *right*, the annotated spectrum of the selected PSM is displayed with quality control plots. Note that it is possible to customize the spectrum annotation using the menu *under* the plot. At the *bottom*, the sequence of the protein is displayed where parts of the sequence presenting no proteolytic peptides are represented by a *thin line* and identified peptides are highlighted in *color* depending on their validation status. The selected peptide is highlighted in *blue*. Note that post-translational modifications (PTMs) are color-coded throughout the interface. Here, one can distinguish phosphorylation and oxidation in *red* and *blue*, respectively, visible in the peptide and PSM tables, as well as under the protein sequence at the *bottom*

2. In the project creation dialog displayed in Fig. 8, select the PeptideShaker output via the *Browse* button next to *Project File* in the *Files Selection* panel. Note that the other fields are filled automatically using the output of PeptideShaker. If it is not the case, fill these manually.
3. Make sure that the reporter ion method selected is *iTRAQ 4-plex*, and click on the cogwheel next to the drop-down menu to open the *Methods Settings* dialog displayed in Fig. 9.
4. Leave the purity coefficients to default (*see Note 18*) and click *OK*.
5. Leave the sample assignment table and quantification settings to default (*see Notes 19 and 20*), and click *Start Quantifying!* to launch the project creation.
6. Upon creation of the project, the results are displayed as shown in Fig. 10. Export the *Default Peptide Report* using the *Export* → *Quantification Features* menu.
7. Repeat the procedure for every PeptideShaker project.

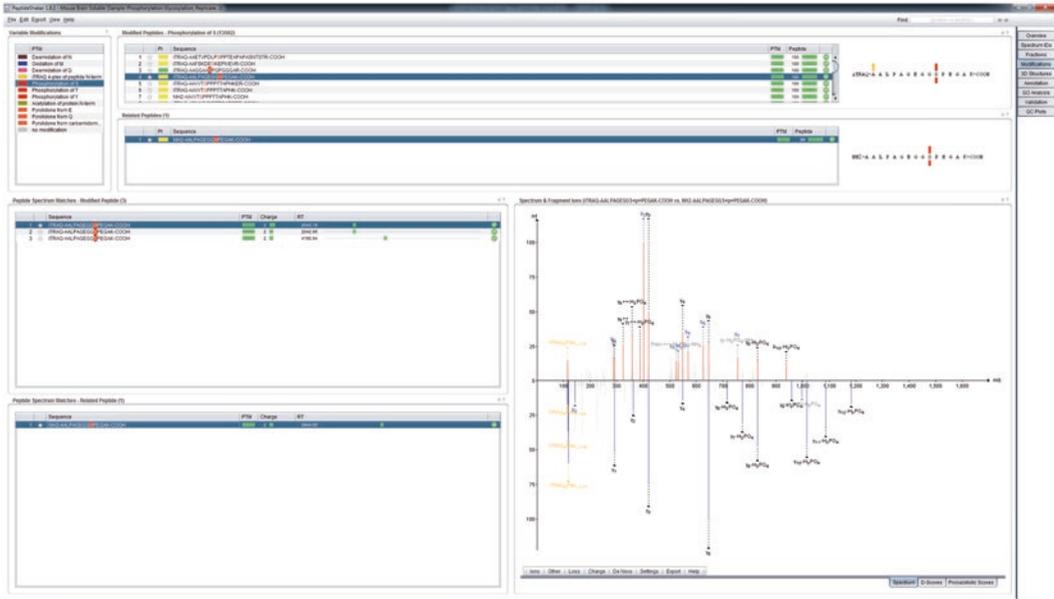


Fig. 7 PeptideShaker Modifications tab. The *Modifications* tab of PeptideShaker lists the peptides carrying the different variable modifications. The modification of interest can be selected in the *top left* table; all peptides carrying this modification will be listed in the *Modified Peptides* table next to it. When a peptide is selected in the *Modified Peptides* table, related peptides are displayed in the *Related Peptides* table, where related peptides are defined as peptides presenting the same sequence but different modifications or cleavage statuses. Here, the selected peptide is found once with an iTRAQ modification on the N-terminus, once without. At the *bottom left*, the different PSMs found for the selected peptides are listed in respective tables, and the spectrum of the selected peptide is displayed to the *right*. Note that if multiple peptides are selected, both spectra will be displayed in a mirrored view, as exemplified here, where one can see that the presence of an extra mass tag on the N-terminus increases the detectability of b ions and reduces the prevalence of y ions. At the *bottom* of the table, the *D*-score and probabilistic scores used to score the localization of PTMs can be visualized in tables. The scores of every PSM aggregated at the peptide level can also be visualized on the peptide sequence next to the peptide tables

3.4 Statistical Analysis

The following analysis is carried out by preparing and running R scripts. The main script *QuickAnalysis.R* contains comments that assist editing and adapting the script to specific projects:

1. Create a folder where to store script and data files. Download R script files from the repository <https://bitbucket.org/veit-veit/ptmanalysis/downloads> and unzip them into the newly created folder. Create subfolders for the different replicates named *Replicate Nr. Enrichment method*, such as *2 phospho*, and copy the respective files from Reporter into the folders. Different folder and file names can be used upon editing of the script (see **Note 21**).
2. Install and start RStudio. Open the R script file *QuickAnalysis.R* (*File* → *Open*). Install the missing libraries by removing the # from the respective lines (see **Note 22**). Execute these lines using either the *Run* button on the upper right or *Ctrl-Enter* (see Fig. 11).

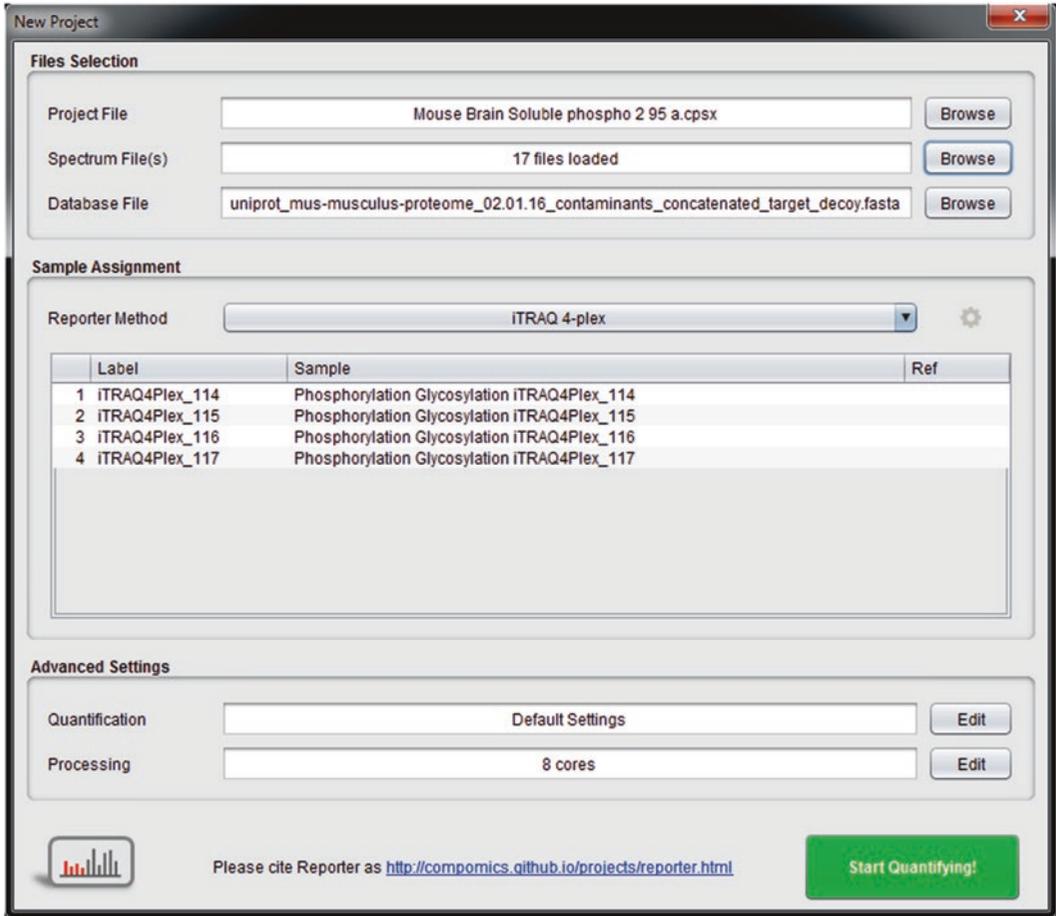


Fig. 8 Reporter project creation dialog. In the *Files Selection* panel at the *top*, it is possible to select the PeptideShaker project used to create the identification project and the spectrum and database files. *Below*, in the *Sample Assignment* panel, it is possible to select the reporter quantification method to use and set the purity coefficients by clicking the cogwheel to the *right*. In the table, it is possible to assign names to the different labels and select reference channels to use for the normalization. In the *Advanced Settings*, it is possible to edit advanced quantification and processing settings. Clicking *Start Quantifying!* at the *bottom* starts the quantification process

3. Set the Working Directory to the folder with the downloaded scripts by *Session* → *Set Working Directory* → *To Source File Location*.
4. Set the main parameters *NumReps*, *NumCond*, and *Ref* to the number of replicates, the number of different experimental conditions, and the reference condition (*see Note 23*) in

```
##### General parameters #####
```

```
NumReps <- 4 # Number of replicates (iTRAQ samples)
```

```
NumCond <- 4 # Number of experimental conditions
```

```
Ref <- 1 # Reference condition for statistical tests.
```

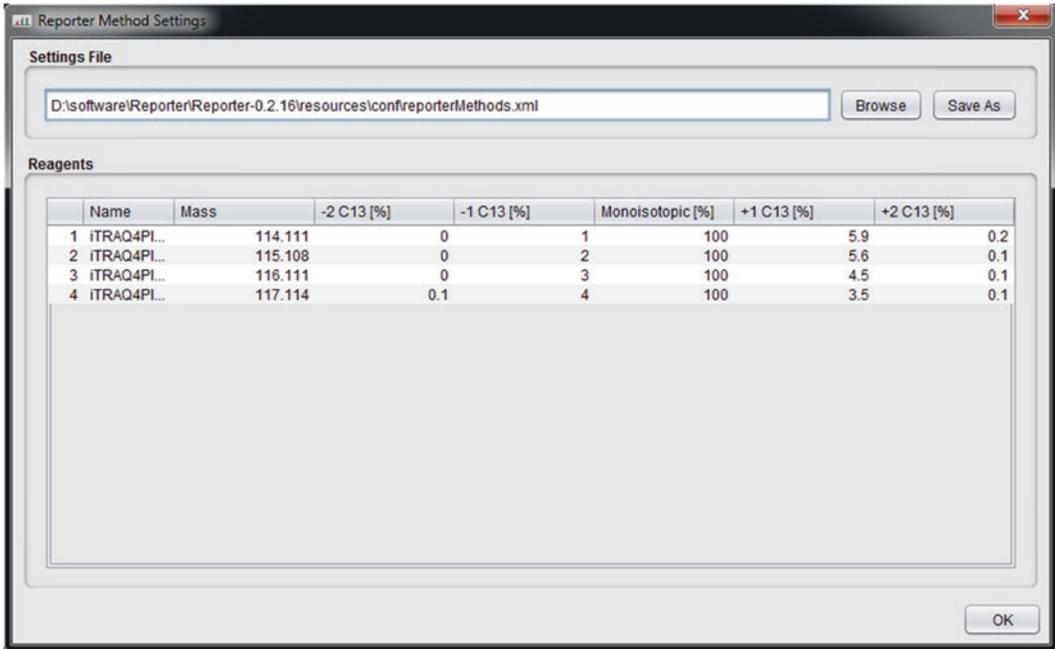


Fig. 9 Reporter Methods Settings dialog. In the *Settings File* panel at the *top*, it is possible to save and open methods files corresponding to the kit used for the labeling. In the *Reagents* table *below*, it is possible to edit the purity coefficients as given by the manufacturer

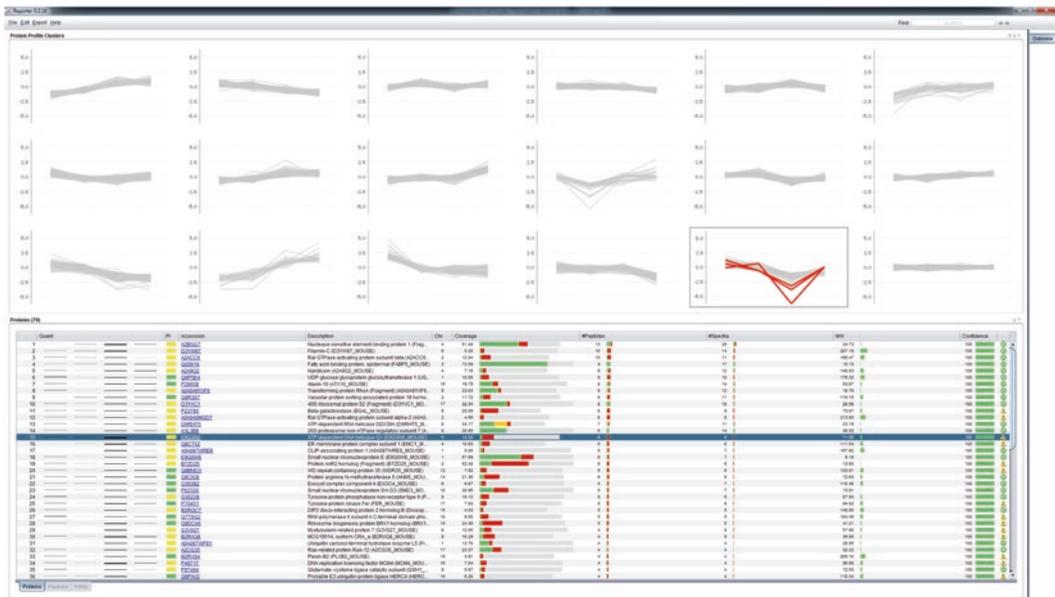


Fig. 10 Reporter Overview tab. The *Overview* tab of Reporter presents the quantification results with clustered regulation profiles at the *top*. Selecting one cluster lists the proteins in the cluster at the *bottom*. The different sample ratios are represented there using sparklines [38], making it easy to find proteins of interest. The proteins selected in the table are in turn colored in *red* in the cluster

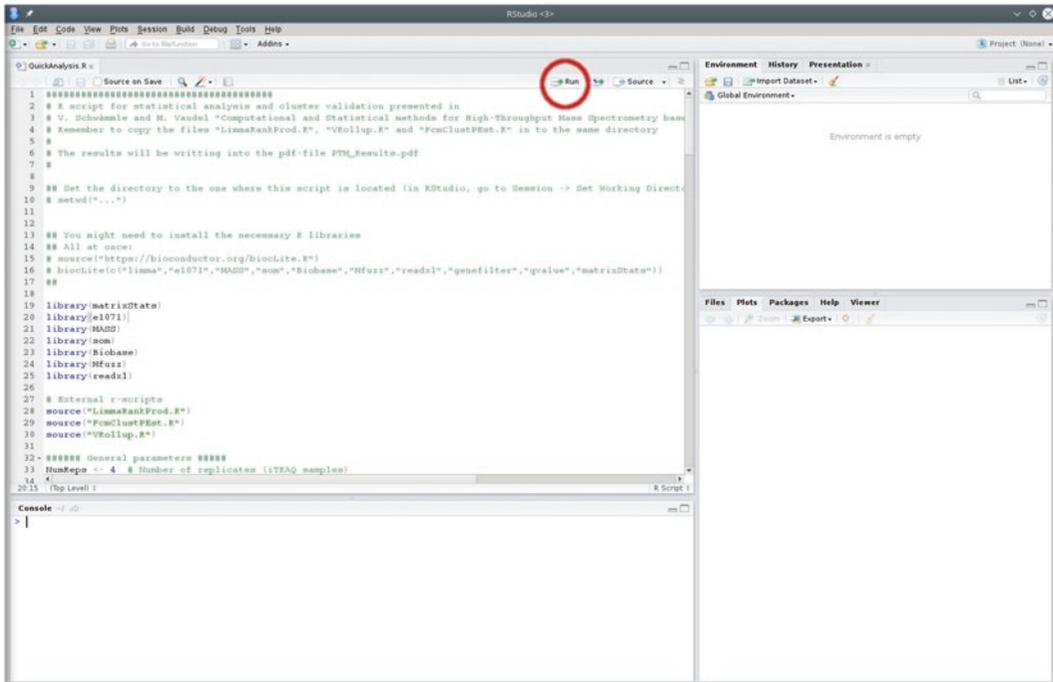


Fig. 11 Run R script QuickAnalysis.R in RStudio. After changing the relevant parameters, the script can be run by selecting the entire code (*ctrl-a*) and left mouse click on the button marked by the red circle. Parts of the script are run by selection of the lines and mouse click on the same button

5. Files will be read from the three folder types *n unmodified*, *n acet*, and *n glyco phospho*, where *n* denotes the replicate number. For changing the file-naming system, see **Note 21**. The script assumes samples to be labeled by iTRAQ. For changing the label to TMT, see **Note 24**. Peptides containing oxidations and carbamidomethylations are filtered out by default.
6. Protein inference is performed by calculating the median of all values that pass a Grubbs' test. The method was adopted from the RRollup function in [39]. It requires a minimum of two unique peptides per protein. For changing this number, see **Note 25**.
7. The data from different replicates is automatically merged. The script outputs the table *Proteins.csv* containing all quantitative protein values and the table *ModifiedPeptides.csv* with all quantifications of the modified peptides. For file name changes, see **Note 26**.
8. The different statistical tests (paired tests within replicates) are performed using the approach presented in [37] that is based on combining moderated *t*-test [40] and rank products [41]. The output is provided as false discovery rates given by *q*-values where correction for multiple testing was carried out using the method of [42]. Output files are *Proteins_qvalues.csv* and *ModifiedPeptides_qvalues.csv* containing *q*-values for the standard *t*-test, moderated

t -test, and rank products, as well as averaged log₂-ratios. For extraction of differentially regulated proteins and modified peptides, filter for features with at least one of the q -values from moderated t -test and rank products below an appropriately set threshold, as, for example, 0.05 yielding a false discovery rate of 5 %. For file name changes, see **Note 27**.

- After setting the right parameters, run the script by selecting the entire code by *ctrl-a* and press *ctrl-enter* or click on the button indicated in Fig. 11. Boxplots of all peptide log₂-ratios for each replicate (Fig. 12), boxplots of merged protein and

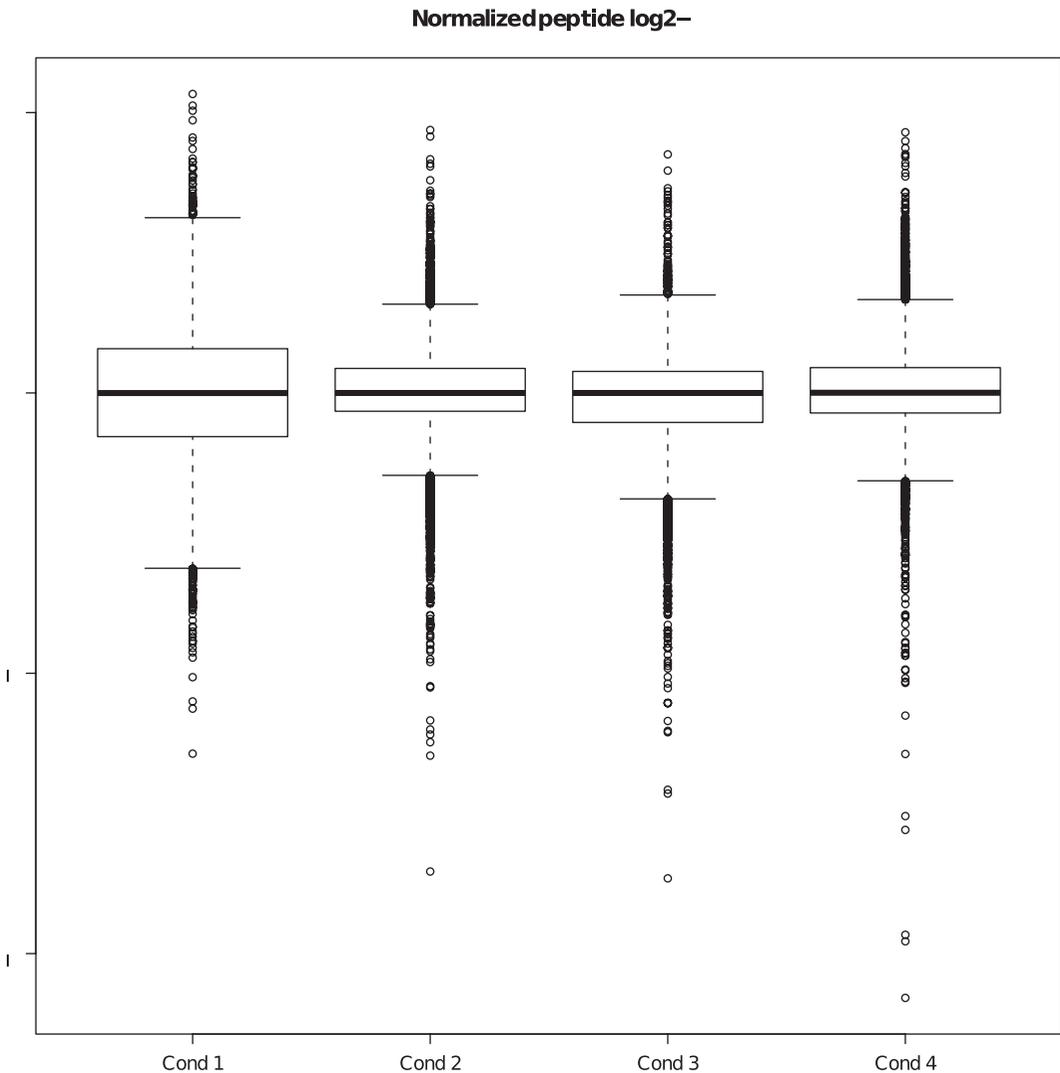


Fig. 12 Boxplot of peptide log₂-ratios from R script QuickAnalysis.R. Figures are written into a pdf file with default name *PTM_Results_Stat.pdf*. For each replicate, the script generates a separate boxplot. The figure shows the distribution of peptide log₂-ratios for the first replicate

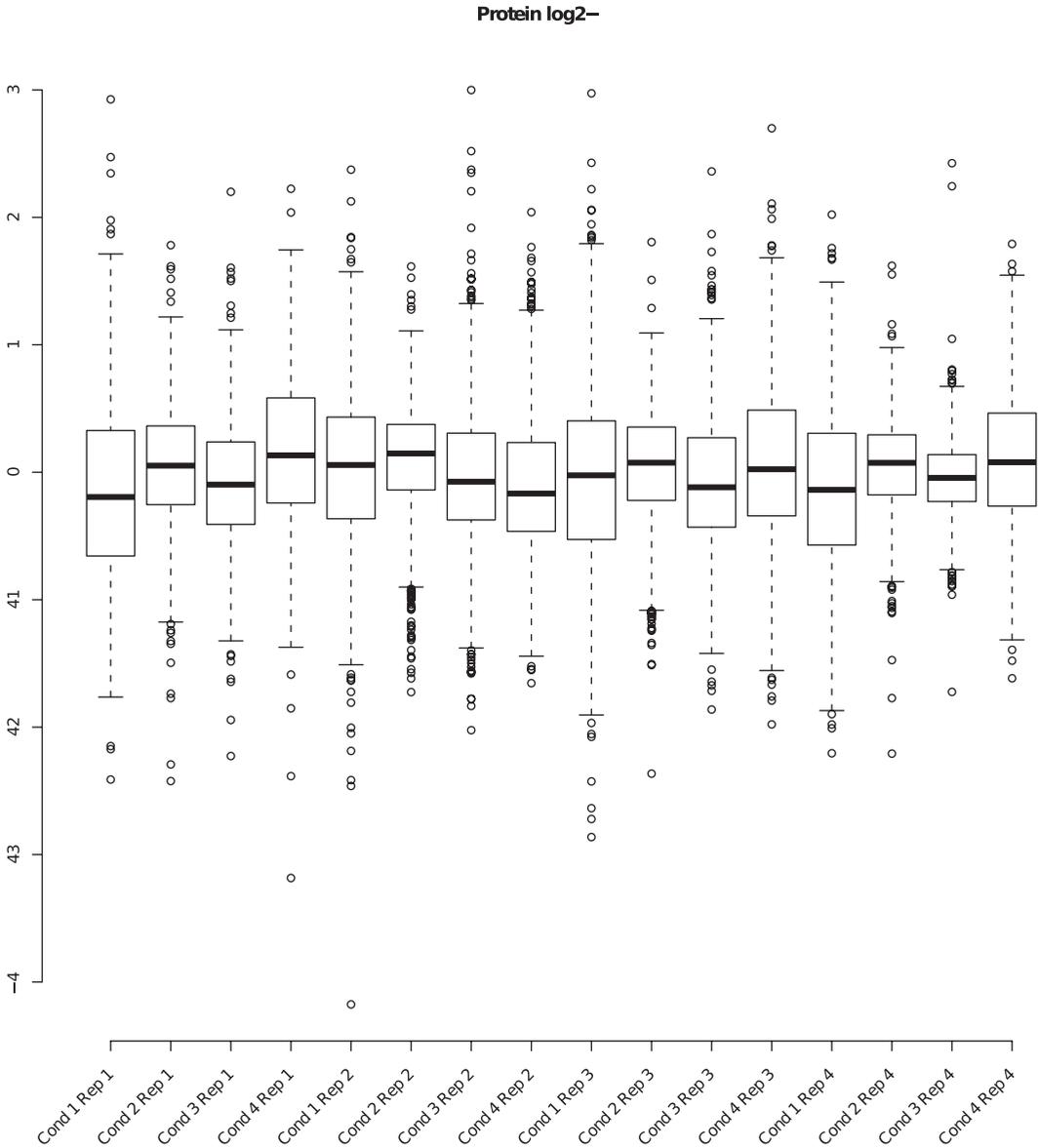


Fig. 13 Boxplot of protein log₂-ratios from R script QuickAnalysis.R. Figure from pdf file with default name *PTM_Results_Stat.pdf*. The plot shows protein log₂-ratios for the entire dataset

modified peptide log₂-ratios (Figs. 13 and 14), *p*-value distributions, and volcano plots for each comparison (Fig. 15) are written into *PTM_Results_Stat.pdf*.

4 Notes

1. UniProt regularly releases new versions of its databases; it is thus important to document the version used in every project and mention it in every scientific work.

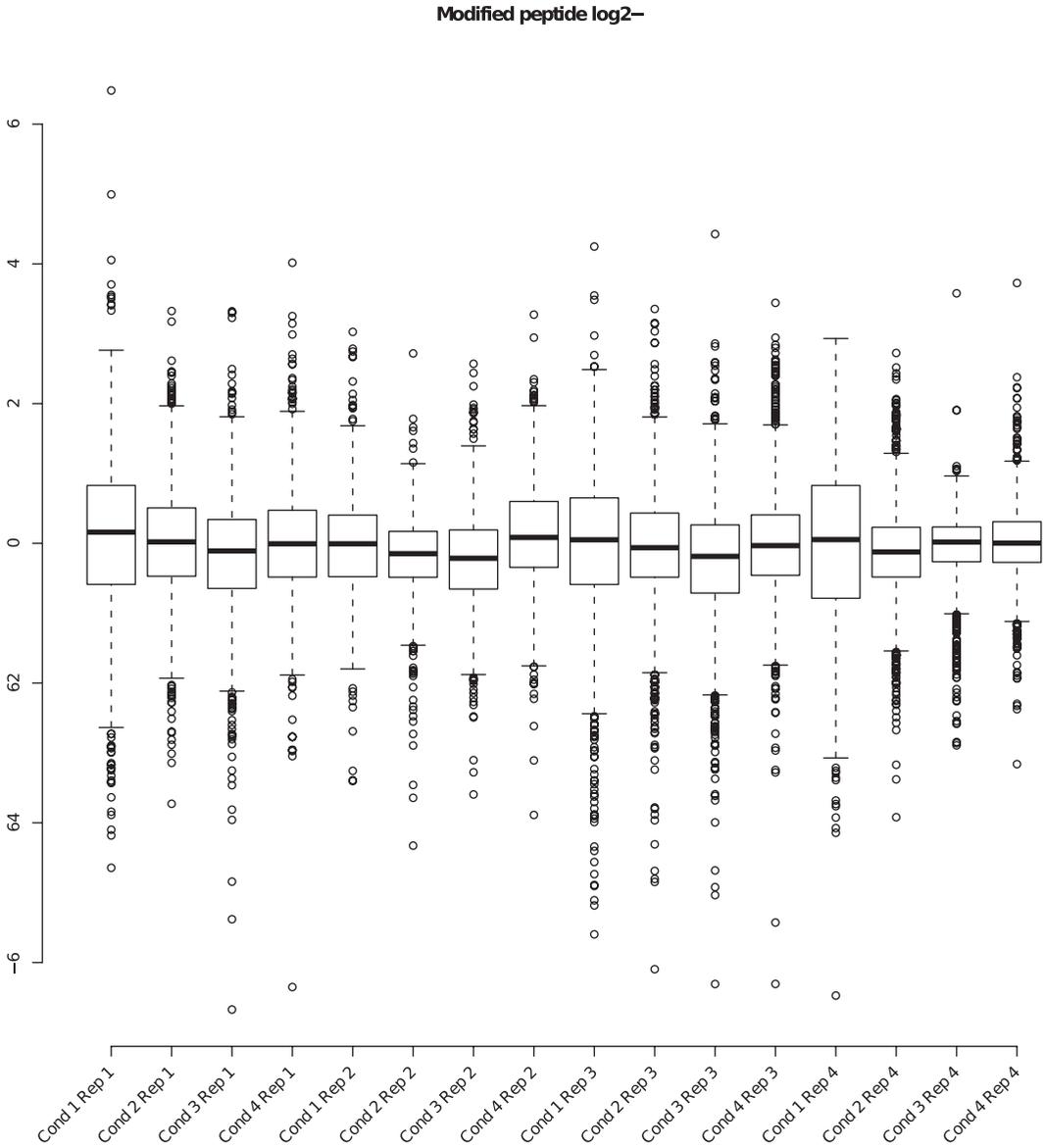


Fig. 14 Boxplot of modified peptide log₂-ratios from R script QuickAnalysis.R. Figure from pdf file with default name *PTM_Results_Stat.pdf*. The plot shows modified peptide log₂-ratios for the entire dataset

2. Spectra originating from contaminants not included in the database can result in the identification of false-positive sequences eventually of nonrandom nature which are poorly tracked by false discovery rate estimation methods [43, 44]. It is thus important to append possible contaminants to the database of proteins of interest.
3. The performance of the software will increase with the computer power available. Make sure that Java 64 bits is installed on 64-bit computers and that the maximum amount of memory is available to the tools. Avoid the usage of synced

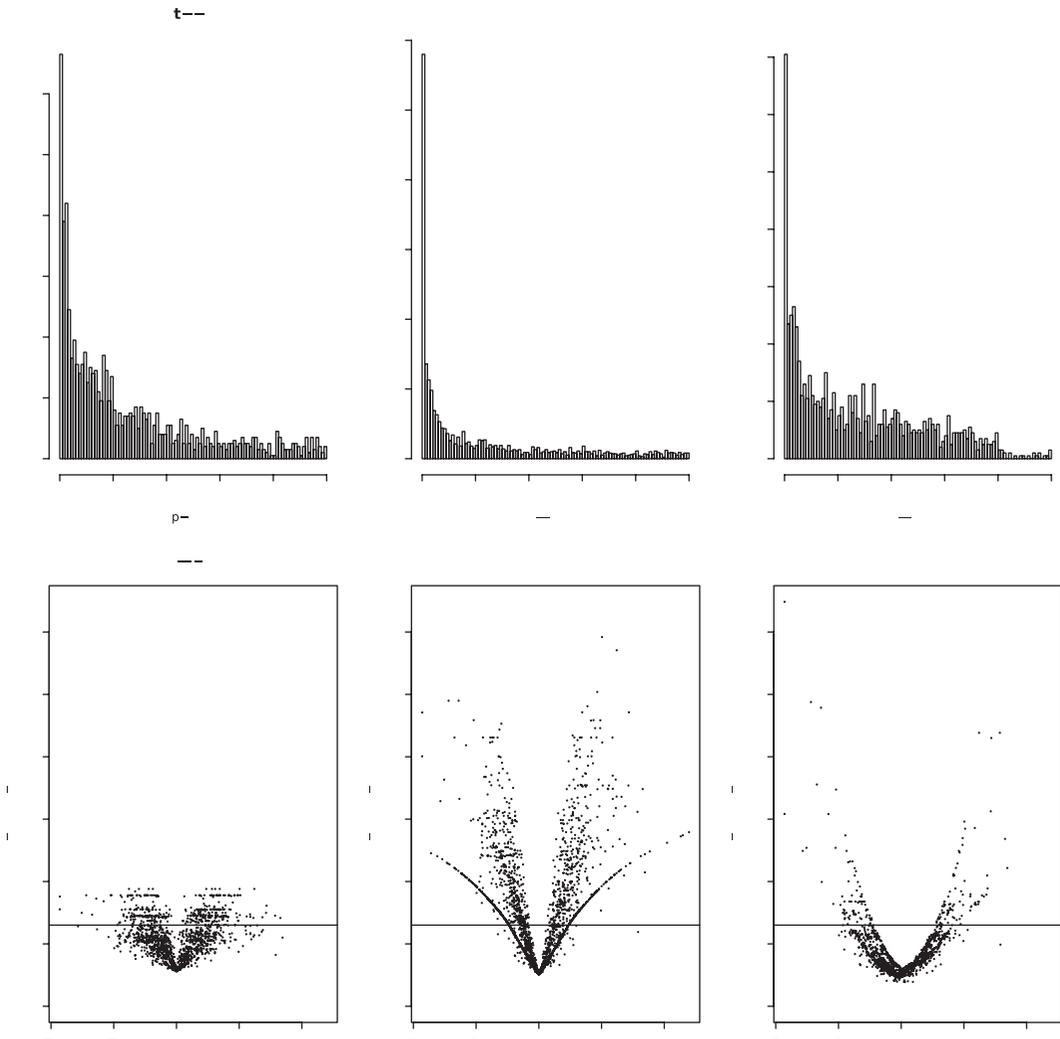


Fig. 15 Statistical plots from R script *QuickAnalysis.R*. Figures are written into a pdf file with default name *PTM_Results_Stat.pdf*. For each comparison, the script generates two separate figures, for proteins and modified peptides. The *upper panels* show p -value distributions for standard t -test, moderated t -test (limma), and rank products. A *peak* at lower p -values denotes the presence of differentially regulated features. These values are transformed into false discovery rates (q -values) and shown as *volcano plots* in the *lower panels*. The *horizontal line* corresponds to a q -value threshold of 0.05. Points *above* this line are found to be differentially regulated with a false discovery rate of at least 5 %

drives and favor solid-state disks (SSD). Make sure that the tools and files are on the same system to avoid read/write limitations. Finally, make sure that you have enough privileges to run the tools. For troubleshooting, help and bug reports and consult the tools' webpages and mailing lists.

4. In order to convert raw files to open formats, msconvert makes use of libraries provided by vendors. These are often only available for Microsoft Windows, and the conversion of raw files must thus be conducted on this operating system.

5. At time of writing, Reporter is released as beta, meaning that while operational, its large-scale usability might require further development. Notably, no command line interface is currently available. For issue reports and feature requests, please refer to the tool webpage (github.com/compomics/reporter/issues).
6. The first time it is used from SearchGUI, a dialog appears allowing the setting of the installation folder of ProteoWizard.
7. If the data requires additional preprocessing, it is possible to set advanced processing options by clicking on the msconvert cogwheel. The different options can be found on the ProteoWizard webpage (proteowizard.sourceforge.net/tools/filters.html).
8. The search settings can influence the identification rate substantially; it is thus recommended to optimize them on few files prior to running the search [45, 46].
9. It is important to give meaningful names to the search settings for later reuse. The different search settings can be managed from the *Edit* → *Identification Settings* menu of the SearchGUI mainframe.
10. Several scores have been developed for PTM localization and are of high importance in PTM studies [47]. These can be selected using the *Localization* settings. The *A*-score [48] and PhosphoRS [49] are readily implemented in PeptideShaker. PhosphoRS is chosen by default for its ability to score multiple modified peptides. In addition, the *D*-score is computed based on the differential PEPs of peptide candidates [50]. The scores are calculated for all variable modifications of the validated peptides. The validation settings can be edited under *Validation Levels*.
11. Depending on the labeling efficiency, it might be possible to have *iTRAQ 4-plex of peptide N-term* as fixed and *iTRAQ 4-plex of Y* as variable.
12. Multiple variable modifications will extend the search time and increase the prevalence of false positives. It is consequently recommended to reduce the number of variable modifications. For this, you can search few fractions with many potential modifications and select the ones actually found in the *Modification Efficiency* plot of the *QC Plots* tab in PeptideShaker.
13. Mass over charge tolerances can be set absolute or relative, the latter provided in ppm. Relative tolerance is mainly used for high-resolution data. Instruments providing high-resolution MS2 spectra are quite recent. Consequently, many search engines do not support MS2 tolerances in ppm. Then, SearchGUI converts the given value using a reference mass editable via the *Edit* → *Advanced Settings* menu. Absolute tolerances are generally referred to using the mass unit Dalton or the *m/z* unit Thomson in the field of mass spectrometry.

14. It is possible to select different packaging options for the search engine results via the *Edit* → *Advanced Settings* menu.
15. Most search engines have advanced settings allowing a better optimization of the search. Some like MS-GF+ notably have specific models dedicated to PTM analyses like phosphorylation studies.
16. From the *Welcome* dialog of PeptideShaker, it is possible to start the *PRIDE Reshake* mode, which allows reprocessing public data from the PRIDE repository. It is hence possible to mine public data for PTMs not originally included in the search.
17. It is possible to export the PeptideShaker project as self-contained zip file for exchange across computers via the *Export* → *PeptideShaker Project As* → *Zip File* menu.
18. A product information sheet or certificate of analysis is provided with every labeling kit containing purity coefficients. These coefficients are kit specific and need to be used in the data interpretation procedure. It is recommended to save these documents and attach them to every publication. For this Chapter, we will use default values.
19. In the *Sample Assignment* table, it is possible to select reference channels. It is recommended to select the samples of lowest variability as reference channels; *see* [46] for more details.
20. In the *Quantification Settings*, it is possible to edit the reporter ion selection settings, the ratio estimation method, and the normalization procedure. There, it is notably possible to edit inclusion and exclusion lists for the normalization, hence avoiding the influence of contaminants.
21. The R command *paste()* generates folder and file names by joining the replicate number with the name of the enrichment methods (here *unenriched*, *glyco phospho*, and *acet*). File names are *Default Peptide Report.xls* by default. For other folder and file names, change the respective part in the *paste(...)* function:

```
filename <- paste(i,"unenriched/Default Peptide Phosphorylation Report.xls")
```

```
filename <- paste(i,"glyco phospho/Default Peptide Phosphorylation Report.xls")
```

```
filename <- paste(i,"acet/Default Peptide Phosphorylation Report.xls")
```

Files can also be read as csv files. Then, change *read_excel* to *read_csv*.

For adding more input files (e.g., enrichment for *ubiquitination*), change

```
File_In <- rbind(File_In,File_In2,File_In3)
```

```

to
filename <- paste(i,"ubiquitination/Default Peptide
Phosphorylation Report.xls")
if (file.exists(filename)) {
File_In4 <- read_excel(filename)
Pos_RR <- which(names(File_In4) %in% sel_columns)
Pos_RR <- c(Pos_RR,Pos_RR[5]+1:(NumCond-1))
File_In4 <- File_In4[,Pos_RR]
}

```

```
File_In <- rbind(File_In,File_In2,File_In3, File_In4)
```

22. Additional R libraries need to be installed the first time the script is run. Remove the # from the following lines:

```

#source("https://bioconductor.org/biocLite.R")
#BiocLite(c("limma","MASS","Biobase","readxl","genefilter",
"qvalue","matrixStats"))

```

After successful execution, add the commenting character # again to avoid unnecessary reinstallations.

23. The statistical tests compare all conditions to the reference condition. Generally, the reference condition corresponds to the control samples or the first time point in a timeline. Depending on the project design, this number might be changed to another condition number than condition 1. For multiple comparisons with different reference conditions, run the script multiple times, each time with the respective reference.

24. From all peptide sequences containing modification information, iTRAQ modifications are removed by

```
File_In[,2] <- gsub("iTRAQ","",File_In[,2])
```

Substitute *iTRAQ* by *TMT* for TMT-labeled data.

25. The line

```
Proteins[[i]] <- VRollup(tmp[,2:ncol(tmp)], tmp[,1], minPep = 2)
```

calls the function located in the file *VRollup.R*. For a different number, e.g., $n=3$, of required unique peptides, change $minPep = 2$ to $minPep = 3$.

26. The script lines

```
write.csv(Prot,"Proteins.csv")
```

```
write.csv(cbind(NMods, accs=Seq2Prot[rownames(NMods)]),
"ModifiedPeptides.csv")
```

can be changed to write the data to differently named files.

27. The script lines

```
write.csv(cbind(qvalModOut,accs=Seq2Prot[rownames(qvalModOut)]),
"ModifiedPeptides_qvalues.csv")
```

```
write.csv(qvalModOut,"Proteins_qvalues.csv")
```

can be changed to write the data to files named differently from default *ModifiedPeptides_qvalues.csv* and *Proteins_qvalues.csv*.

Acknowledgments

VS was funded by the Danish Council for Independent Research and the EU ELIXIR consortium (Danish ELIXIR node). This work was conducted as part of the EuPA Bioinformatics Community (EuBIC) initiative supported by the European Proteomics Association (EuPA).

References

1. Minguez P, Letunic I, Parca L et al (2013) PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic Acids Res* 41:D306–D311
2. Hunter T (2000) Signaling—2000 and beyond. *Cell* 100:113–127
3. Munoz J, Heck AJ (2014) From the human genome to the human proteome. *Angewandte Chem Int Ed Engl* 53:10864–10866
4. Altelaar AF, Munoz J, Heck AJ (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet* 14:35–48
5. Solari FA, Dell’Aica M, Sickmann A et al (2015) Why phosphoproteomics is still a challenge. *Mol Biosyst* 11(6):1487–1493
6. Olsen JV, Mann M (2013) Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics* 12:3444–3452
7. Tran JC, Zamdborg L, Ahlf DR et al (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480:254–258
8. Chait BT (2006) Chemistry. Mass spectrometry: bottom-up or top-down? *Science* 314:65–66
9. Perkel JM (2015) Top-down proteomics: turning protein mass spec upside-down. *Science* 349:1243–1245
10. Gevaert K, Van Damme P, Ghesquiere B et al (2007) A la carte proteomics with an emphasis on gel-free techniques. *Proteomics* 7:2698–2718
11. Vaudel M, Barsnes H, Bjerkvig R et al (2016) Practical considerations for omics experiments in biomedical sciences. *Curr Pharm Biotechnol* 17:105–114
12. Schwämmle V, Verano-Braga T, Roepstorff P (2015) Computational and statistical methods for high-throughput analysis of post-translational modifications of proteins. *J Proteomics* 129:3–15
13. Bantscheff M, Schirle M, Sweetman G et al (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 389:1017–1031
14. Bantscheff M, Lemeer S, Savitski MM et al (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* 404:939–965
15. Geiger T, Cox J, Ostasiewicz P et al (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* 7:383–385
16. McAlister GC, Huttlin EL, Haas W et al (2012) Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem* 84:7469–7478
17. Ross PL, Huang YN, Marchese JN et al (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 3:1154–1169
18. Thompson A, Schafer J, Kuhn K et al (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 75:1895–1904
19. Vaudel M, Sickmann A, Martens L (2010) Peptide and protein quantification: a map of the minefield. *Proteomics* 10:650–670
20. Edwards AV, Edwards GJ, Schwämmle V et al (2014) Spatial and temporal effects in protein post-translational modification distributions in the developing mouse brain. *J Proteome Res* 13:260–267

21. Edwards AV, Schwämmle V, Larsen MR (2014) Neuronal process structure and growth proteins are targets of heavy PTM regulation during brain development. *J Proteomics* 101:77–87
22. Martens L, Hermjakob H, Jones P et al (2005) PRIDE: the proteomics identifications database. *Proteomics* 5:3537–3545
23. Vizcaino JA, Deutsch EW, Wang R et al (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 32:223–226
24. Vaudel M, Venne AS, Berven FS et al (2014) Shedding light on black boxes in protein identification. *Proteomics* 14:1001–1005
25. Kessner D, Chambers M, Burke R et al (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24:2534–2536
26. French WR, Zimmerman LJ, Schilling B et al (2015) Wavelet-based peak detection and a new charge inference procedure for MS/MS implemented in ProteoWizard's msConvert. *J Proteome Res* 14:1299–1307
27. Vaudel M, Barsnes H, Berven FS et al (2011) SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 11:996–999
28. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20:1466–1467
29. Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6:654–661
30. Dorfer V, Pichler P, Stranzl T et al (2014) MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res* 13:3679–3684
31. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 5:5277
32. Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 3:1234–1242
33. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics* 13:22–24
34. Diament BJ, Noble WS (2011) Faster SEQUEST searching for peptide identification from tandem mass spectra. *J Proteome Res* 10:3871–3879
35. Cox J, Neuhauser N, Michalski A et al (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10:1794–1805
36. Vaudel M, Burkhart JM, Zahedi RP et al (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 33:22–24
37. Schwämmle V, Leon IR, Jensen ON (2013) Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates. *J Proteome Res* 12:3874–3883
38. Barsnes H, Vaudel M, Martens L (2015) JSparklines: making tabular proteomics data come alive. *Proteomics* 15:1428–1431
39. Polpitiya AD, Qian W-J, Jaitly N et al (2008) DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* 24:1556–1558
40. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3
41. Breitling R, Armengaud P, Amtmann A et al (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* 573:83–92
42. Storey JD (2002) A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol* 64:479–498
43. Colaert N, Degroev S, Helsens K et al (2011) Analysis of the resolution limitations of peptide identification algorithms. *J Proteome Res* 10:5555–5561
44. Knudsen GM, Chalkley RJ (2011) The effect of using an inappropriate protein database for proteomic data analysis. *PLoS One* 6:e20873
45. Muth T, Kolmeder CA, Salojarvi J et al (2015) Navigating through metaproteomics data: a logbook of database searching. *Proteomics* 15:3439–3453
46. Vaudel M, Sickmann A, Martens L (2014) Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics. *Biochim Biophys Acta* 1844:12–20
47. Chalkley RJ, Clauser KR (2012) Modification site localization scoring: strategies and performance. *Mol Cell Proteomics* 11:3–14
48. Beausoleil SA, Villen J, Gerber SA et al (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 24:1285–1292
49. Taus T, Kocher T, Pichler P et al (2011) Universal and confident phosphorylation site localization using phosphoRS. *J Proteome Res* 10:5354–5362
50. Vaudel M, Breiter D, Beck F et al (2013) D-score: a search engine independent MD-score. *Proteomics* 13:1036–1041

Cross-Species PTM Mapping from Phosphoproteomic Data

Rima Chaudhuri and Jean Yee Hwa Yang

Abstract

Protein post-translational modifications (PTMs) are crucial for signal transduction in cells. In order to understand key cell signaling events, identification of functionally important PTMs, which are more likely to be evolutionarily conserved, is necessary. In recent times, high-throughput mass spectrometry (MS) has made quantitative datasets in diverse species readily available, which has led to a growing need for tools to facilitate cross-species comparison of PTM data. Cross-species comparison of PTM sites is difficult since they often lie in structurally disordered protein domains. Current tools that address this can only map known PTMs between species based on previously annotated orthologous phosphosites and do not enable cross-species mapping of newly identified modification sites. Here, we describe an automated web-based tool, PhosphOrtholog, that accurately maps annotated and novel orthologous PTM sites from high-throughput MS-based experimental data obtained from different species without relying on existing PTM databases. Identification of conserved PTMs across species from large-scale experimental data increases our knowledgebase of evolutionarily conserved and functional PTM sites that influence most biological processes. In this Chapter, we illustrate with examples how to use PhosphOrtholog to map novel PTM sites from cross-species MS-based phosphoproteomics data.

Key words Cross-species, Post-translational modification (PTM), PTM site mapping, Web application

1 Introduction

Protein post-translational modifications (PTMs) such as phosphorylation regulate cellular processes creating complex signaling networks that facilitate cell development along with inter- and intracellular responses. PTMs are widely recognized as key regulators of multiple cellular processes, including metabolism. Defects in PTMs have been extensively linked with human diseases, implying their importance in maintaining normal cellular functions [1]. It is estimated that approximately 1 million different residues can be phosphorylated by around 518 protein kinases in humans under specific conditions [2]. Currently, only ~21 % (20,9958 nonredundant human phosphosites) of these possible sites are annotated in PhosphoSitePlus [3–5]. Furthermore, in model organism such as

rat and fly, there is an even sparser repertoire of phosphorylation sites found due to the lack of extensive large-scale phosphoproteome studies.

Identifying novel PTM sites in different organisms has expedited our understanding of key signal transduction mechanisms in cellular biology. Landry and colleagues found that functional PTMs are more likely to be evolutionarily conserved across humans and model organisms such as mice, rats, and flies [6]. This has led to the frequent use of species conservation as a criterion for selecting specific phosphorylation sites of interest for functional characterization. Hence, an easy to use tool facilitating the mapping of PTMs across different target species is of particular benefit to the proteomics community.

1.1 Challenges in Mapping Cross- Species Proteomics Data

It is nontrivial to integrate and accurately map PTM sites across multiple species in a high-throughput manner. Mapping novel PTM sites identified from new proteomics experiments by the scientists directly without advanced bioinformatics support can be a daunting task. The current methodological challenges and shortcomings in the area of cross-species mapping are listed below.

1. Modification sites often exist in unstructured flexible domains of proteins in the least conserved domains across species, making mapping on a large scale a substantial challenge.
2. Differences exist between the protein sequence database versions used while annotating peptides to proteins. This can occur in situations where we integrate datasets from different species as well as data across multiple experiments from the same species.
3. Conservation of sequence positions is particularly poor in distantly related organisms such as yeast [7]. While the exact amino acid position of a modified residue in the primary sequence may in some cases be conserved between mammalian species (i.e., mouse, rat, and human), it is nontrivial to directly compare PTM sites that differ in their amino acid positions. For example, the functionally relevant phosphorylation of protein acetyl-CoA carboxylase 1 (ACACA) is at S344 in humans, S343 in mouse, and S285 in yeast; hence, they cannot be directly mapped to each other as orthologous sites without aligning their sequences.
4. Existing databases such as PhosphoSitePlus (PSP) database only contain modification sites identified experimentally in the target species rather than the full repertoire of orthologous residues, thereby impeding seamless discovery of novel orthologous sites. For example, in the case of Unc-51-like kinase 1 (ULK1), the functionally relevant phosphorylation at S758 in human and S757 in mouse that has been experimentally identified is annotated in PSP with a unique SITE_GROUP_ID. At

the same time, the homologous site in rat has not been identified in any proteomics experiments and is therefore absent. In this situation, comparison of rat phosphoproteomics experiments with another species using PSP would exclude this biologically important phosphorylation site, despite its presence in this species. One possible solution for mapping a single site is to engage in manual site-by-site query of PTMs sites across different species. Unfortunately, performing such mapping tasks in a systematic and high-throughput manner for large datasets is currently not possible.

1.2 Cross-Species Mapping Resources

Resources surrounding phosphoproteomics data mainly consist of PTM repositories and phosphorylation prediction tools.

PTM data repositories: Currently available PTM repositories such as PSP [3–5], PhosphoBlast [8], Phospho.ELM, and PHOSIDA [9, 10] are repositories of known PTMs and not high-throughput data mapping tools. They can only be used to obtain information about conservation of *known* modification sites across species. These databases do not allow unannotated phosphorylation sites that differ in position between species to be mapped to each other readily.

Phosphorylation prediction tool: DAPPLE [11] which is “a pipeline for the homology-based prediction of phosphorylation sites” aims to identify putative sites in an organism of interest by using known phosphorylation sites from other organisms. Although DAPPLE [11] can handle high-throughput input data from one species and predict phosphorylation sites in a target species of interest, it uses experimental evidence of only known phosphorylation sites obtained by searching through the abovementioned databases; hence, mapping of novel PTM sites from the user’s own data cannot be achieved using DAPPLE [11]. The above resources rely solely on already identified orthologous sites curated from literature.

More recently, an automated web-based tool, PhosphOrtholog (www.phosphortholog.com), was developed [12] which allows batch mapping of large species-specific PTM datasets to estimate the overlap between them at a site-specific level. Through the use of PhosphOrtholog, both unannotated phosphorylation sites that differ in position between species and known orthologous PTM sites can be easily mapped for cross-species data comparisons. Moreover, PhosphOrtholog is not restricted to use with only phosphorylation data but can be extended to map any type of PTMs such as ubiquitination, methylation, and acetylation, among others. With PhosphOrtholog, users can now find conserved functional PTM sites across multi-species datasets through the click of a few buttons in an easy-to-use web interface.

2 Materials

2.1 Datasets

Mouse phosphoproteomics: Supplementary Table S2 from Zhong J et al. 2015 [13].

Human phosphoproteomics: Supplementary Table S1 from Hoffman and Parker et al. 2016 [14].

The data files must contain columns including the following information in order to prepare the input files for PhosphOrtholog.

- UniProt ID.
- Modification site residue type.
- Modification residue number within the UniProt ID sequence.

2.2 Internet Resources

Internet access to the tool www.phosphortholog.com and databases contained therein such as:

- Ortholog reference databases (e.g., human_mouse, mouse_rat, etc.).
- PhosphoSitePlus database.

3 Methods

In the phosphoproteomics data analysis pipeline, the role of PhosphOrtholog is illustrated through Fig. 1. The pipeline can be divided into four phases.

Phase 1, the process of sample extraction and preparations from the species of interest for the MS-based experiments are performed. The MS experiment generates spectral data which are then preprocessed and evaluated by spectral analysis software such as MaxQuant [15] through multiple steps.

Phase 2, the software generates output which contains protein and modification site level annotations along with quantitative intensity measures from each experiment conducted in phase 1; these are often large spreadsheets with data in the form of multidimensional matrices (PTM sites detected as rows and samples/experimental conditions as columns).

Phase 3, appropriate PTM site annotation columns (UniProt ID, modified amino acid type, and residue number) are parsed out from these output files from each species and merged together in the desired input format for PhosphOrtholog, which will be described in detail below. At this level, any publicly available proteomics data can similarly be parsed for appropriate columns and used as input for PhosphOrtholog (*see Notes 1–5* for input preparation tips).

Phase 4, the PhosphOrtholog web interface is used to map common sites between the two input datasets of different species. The resulting mapped table returns the newly mapped sites associated with a calculated E-value per match, whereas previously

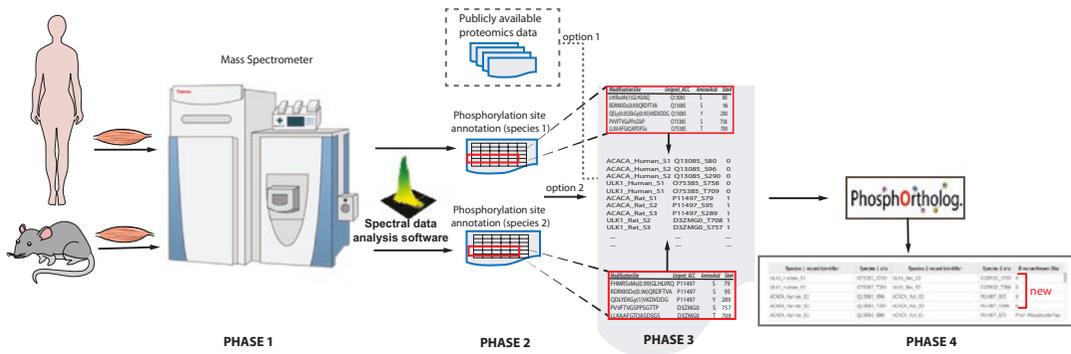


Fig. 1 Role of PhosphOrtholog in the MS-based PTM data analysis pipeline. Illustration of the broadly divided four stages of MS-based PTM experiments, in Phase 1, sample extraction, and preparation tasks are conducted from human and rat muscle tissues for the MS-based phosphoproteomics experiment. Phase 2 marks the raw spectral data analysis to generate peptide and protein annotations along with intensity measures for the PTMs induced by the experimental design in each species. In Phase 3, output from Stage 2 is parsed to extract information from each species to generate the desired input format for PhosphOrtholog-based mapping. In Phase 4, the sites mapped by PhosphOrtholog are obtained, which are either annotated as newly mapped with a calculated E-value (4 out of 5 input sites were not mapped before, identified with E-value of 0) or with “From PhosphoSitePlus” if the mapping was previously known. From Chaudhuri and Sadrieh et al. [12]

known sites are marked by “From PhosphoSitePlus” since PhosphOrtholog scans through the PSP database to extract annotated orthologous site matches.

3.1 Pipeline Algorithm

The central algorithmic component of PhosphOrtholog is the sequence alignment between orthologous protein sequences from different species. This is a deterministic process, where a match is reported only if a match between amino acid type and position occurs between the two aligned sequences. There are no random/predictive elements involved in the generation of the output data from a given input. There is either a match of residue type and modification site number between the two species compared or not. The accuracy of this deterministic mapping is entirely dependent on the pairwise sequence alignment of the sequences, which is reflected by the E-value score representing the statistical significance of each alignment. A reported E-value of 0 is a rounded probability ($< \sim 1e - 250$), which means that there is essentially no chance that alignment could occur by chance.

In the next few sections, we will describe the usage and utility of Phase 3 and Phase 4 of PhosphOrtholog in more detail.

3.2 Input Requirements for PhosphOrtholog

PhosphOrtholog provides a web-based data entry form with three columns as input as shown in Table 1 that can be copy-pasted manually into the browser interface directly or uploaded as a comma-delimited (.csv) file without any headers. It is mandatory to input information for both species in order to complete mapping between them (*see Note 1*). The details for the three input columns are the following:

Table 1
Input format for PhosphOrtholog

Record identifier	Site	Species code
H1	A0AVK6_S608	0
H2	A0FGR8_S676	0
...
H8511	Q9Y6Y0_S322	0
M1	F6S0D5_S33	2
M2	D6RI64_S17	2
...
M5527	Q5SSH7-2_S1538	2

1. **Column 1** should contain a unique identifier or “Record Identifier” for each site in the human/rat/mouse/fly data. This could be any unique series of numbers, letters, and special characters. For example, column 1 in Table 1 shows a simple example of a numerical series denoting the row number for each record appended with h (human) or m (mouse) depending on the species the record originates from. This is only required for “record keeping” or maintaining annotation for each record for downstream analysis and is not used in the mapping algorithm (*see Notes 2 and 3*).
2. **Column 2** must contain protein and site information obtained from proteomics data. Required data format should follow UniProt ID_modified amino acid in one letter code followed by the modification site number with respect to the whole protein (for the UniProt ID provided). For example, Q13085_S23, where Q13085 is the UniProt ID, S is the modified amino acid (Serine), and 23 is the position of the modified residue with respect to the protein Q13085. An underscore sign must separate the UniProt ID and site modification annotation (*see Notes 2 and 3*).
3. **Column 3** must contain a species identifier: 0 for human, 1 for rat, 2 for mouse, and 3 for fly.

3.3 Illustrative Example of Input Preparation for PhosphOrtholog from Proteomics Data

Since manual input preparation could be a potential bottleneck for users, in this section, we will illustrate the process of extracting relevant information from real proteomics datasets to generate an input file for PhosphOrtholog. Spectral data analyzed by processing software such as MaxQuant [15] outputs a plethora of quantitative and qualitative information as columns within a spreadsheet from Phase 2 of the analysis pipeline in Fig. 1.

The **first column** could be any unique identifier for each record as described in the input requirements section. For example, if the user desires to map human and rat PTM site data, each containing x and y records, respectively, the first column for human could range from H1 to Hx and that for rat from R1 to Ry. The **third column** is a 0, 1, 2, or 3 depending on the species.

The **second column** needs more preparation. Typically, the output of spectral data analysis software such as MaxQuant [15] contains protein/peptide annotations and other experimental details as various columns. One of the columns contains the UniProt IDs assigned to a searched peptide against the software's FASTA sequence database version and often termed as "leading proteins" which is usually a string of UniProt IDs such as A0FGR8;A0FGR8-2;A0FGR8-4;A0FGR8-5;A0FGR8-6;F2Z3K9;H7BX11 which is used as an example for a PTM site in protein ESYT2 for the peptide sequence window SAQVKRPSVSKEGRK (from Supplementary Table 1 of Hoffman and Parker et al. 2015). We can also obtain a column with the sequence localization probability which indicates the likelihood of a particular residue (S/Y or T for phosphorylation) being modified within that sequence, for example, the column entry of rPsVSk_S(3): >90 %; S(5): <10 % for ESYT2 would indicate that the Serine at position 3 (marked by S(3)) within the sequence RPSVSK has been identified as a modification site with a probability of greater than 90 %, whereas the Serine at position 5 is not likely to be identified as a modification site since the probability of this position being modified is less than 10 % (*see Note 4*). Users often define a probability threshold by which they define reliable detection of modification sites (a cutoff of 90 % is often appropriate). The amino acid position of the modified serine residue within the UniProt ID is also indicated in one of the column output, for example, within the first UniProt ID A0FGR8 for ESYT2, the modified serine S3 lies at position 676 in context of the entire protein sequence. The residue type (S/Y/T for phosphorylation) is often output as a separate column as well; if not, it can be easily parsed from the localization probability column. Once the abovementioned information is located within the proteomics data file, users should parse out (1) only the first UniProt ID (under the assumption that this is the most reliable protein assignment for the peptide sequence searched by the analysis software), (2) the modified residue number within that UniProt ID sequence, and (3) modified residue type for formatting input to PhosphOrtholog.

Parsing can be achieved using any scripting language such as R, perl, awk, or sed (*see Note 5*). For example, within Microsoft Office Suite, Excel can be used to concatenate various columns by using the "&" symbol. Columns with individual entries of A0FGR8, S, and 676 in say cells A1, B1, and C1 can be concatenated using the command "A1&"_"&B1&C1" in a new column D, which will

result in A0FGR8_S676. This column would serve as the **second column** of PhosphOrtholog input (described in Subheading 3.2) for species 1 from experimental dataset 1. This process should be repeated for species 2 using experimental dataset 2.

For practice, users can download publicly available mouse phosphoproteomics data generated by Zhong et al. 2015 [13] and human data from Parker and Hoffman et al. 2016 [14] as instructed in the Materials section. Once data for both species to be compared is formatted as described above, the individual species data can be appended together into one large three-column sheet of length $x + y$. The resulting input will resemble Table 1 and can be used by PhosphOrtholog for cross-species site mapping; the columns should be comma separated if uploaded as a file.

3.4 Parameter Specification and Site Navigation Within PhosphOrtholog

In Fig. 2, the key features of the input and output of PhosphOrtholog are highlighted. The input data can be directly entered through the user interface via copy-paste functionality into the “Preview of input data set” table or alternatively uploaded as a comma-delimited (.csv) file through the “Upload” button (Fig. 2a). Example input data can be found on the web page that can be directly input into PhosphOrtholog by clicking the “Use above example” button; this will populate the data into the input table ready for mapping. Within the web application, users can also find an additional example dataset obtained from insulin-stimulated phosphoproteomics experiments in human and rat skeletal muscles. This data can be downloaded through the “download” link and used as an example input file by uploading it into the web browser. An example of the required input data format can also be found in Table 1.

Once the input data is uploaded and displayed in the input table, the user should click on the “Map” button and the system returns the results on the screen in the output table below in Step #3 (Fig. 2b) which can be copy-pasted or downloaded as a comma-separated file by clicking the “Download” button.

3.5 Output

Once the mapping is complete, we report three summary counts and a five-column output file. The three summary counts consist of (a) the number of novel sites mapped by PhosphOrtholog in the dataset; (b) the percentage of data that could not have been

Fig. 2 (continued) the example data table and uploaded to the UI through the “Upload” button. **(b)** The “Map” button in “Step 2” starts the mapping task; the progress bar above the output table in “Step 3” tracks the progress of the mapping function. This will give a rough estimate of how long the job will take to finish for large datasets. The first two columns in the mapped output table indicate the species 1 record identifier and PTM site details which is mapped to the orthologous species 2 site information shown in the third and fourth columns. The last column indicates the E-value significance score from the pairwise sequence alignment of the orthologous proteins. If the PTM site is a known mapped site from PhosphoSitePlus database, then this column reports “From PhosphoSitePlus” instead of an E-value. From Chaudhuri and Sadrieh et al. [12]

a

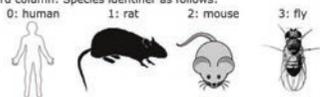
PhosphOrtholog.

A tool to map post-translational modification sites on proteins orthologous across species
Created by Rima Chaudhuri and Arash Sadrieh.

Step #1: Upload input data set

Prepare your input data set in the following format:

- 1st column: Unique record identifier for each site in the human/rat/mouse/fly data set.
- 2nd column: UNIPROT ID, modified amino acid type, and the site position. E.g., "Q93100_S700"
- 3rd column: Species identifier as follows:



For example:

Species Record Identifier	Site	Species Code
QNGSND(0.001)DRYS(0.999)DNEEDSK	P42167_S184	0
RAES(0.996)RT(0.193)S(0.811)VGS(1)QR	Q9BR39_S235	0
S(0.002)ES(0.992)RT(0.005)S(0.001)LGSQR	Q2PS20_S228	1
QNGSND(0.001)DRYS(0.999)DNEEDSKIELK	Q62733_S183	1

Note: If you don't have your own data set, you can try it with one of the following. The first one, a human-rat sample data set (download), contains information for human skeletal muscle insulin-regulated phosphoproteome (1187 human sites quantified; 551 unique protein accessions) and rat skeletal muscle myotube insulin-regulated phosphoproteome (10,033 sites quantified; 3050 unique accessions). Since, this file contains a large number of records, it might take a long time to complete mapping. For a quick test, use this human-mouse sample data set (download) or this human-rat sample data set (download).

Please remember to input modification site information for two different species for PhosphOrtholog to complete mapping between them. Please ensure that the input is in comma separated file format (.csv).

Upload

Use above example

Clear

Input data by either pasting/typing directly into the table below. Otherwise, you can input a CSV or comma-delimited file in the column-format above by selecting the "Upload" button above.

Preview of input data set

Species Record Identifier	Site	Species Code

b

Step #2: Run mapping algorithm

After successfully uploading the input data set, you can now run the PhosphOrtholog mapping algorithm. This will enable mapping between human-mouse, human-rat, human-fly, fly-mouse or rat-mouse phosphorylation sites quantified in independent MS-based phosphoproteomic experiments. The mapping result will be displayed under step #3.

Map

Step #3: Download results data set

You can find the mapping result below. The first two columns represents the human identifier and site information which is mapped to the orthologous sites in the model organism shown in the third and fourth columns. The "E-value/known Site" represents the Novel Site E-value from the pairwise sequence alignment of the orthologous sequences across different species. If the sites are known and reported in PhosphoSitePlus, the 'E-value/known Site' column says "from PhosphositePlus DB". E-value represents the probability of sequence alignment occurring by chance. An E-value of 0 is a rounded probability ($< \sim 1e-250$), which means that there is no chance that alignment could occur by chance. It is a statistical significance calculation based on the quality of alignment (the similarity score based on BLOSUM62 substitution matrix used for alignment in this study) and the size of the database.

Progress bar

Novel sites mapped exclusively by PhosphOrtholog between Human and Rat: 1

% of novel sites mapped in input data by PhosphOrtholog: 50

% of known sites from PhosphoSitePlus also mapped by PhosphOrtholog: 100

Mapping results summary

Species 1 record identifier	Species 1 site	Species 2 record identifier	Species 2 site	E-value/known Site
RAES(0.996)RT(0.193)S(0.811)VGS(1)QR	Q9BR39_S235	S(0.002)ES(0.992)RT(0.005)S(0.001)LGSQR	Q2PS20_S228	0
QNGSND(0.001)DRYS(0.999)DNEEDSK	P42167_S184	QNGSND(0.001)DRYS(0.999)DNEEDSKIELK	Q62733_S183	From PhosphositePlus

Download

Proteomics Resource

Download Ortholog Reference Databases

In the process of mapping UNIPROT IDs between species, we compiled and used the following databases, that can be downloaded:

Fly: Mouse download, Fly: Rat download, Human: Fly download, Human: Mouse download, Human: Rat download and Rat: Mouse download

Fig. 2 Snapshots of the PhosphOrtholog web application. (a) Instructions for generating the input data format, including each column description is shown in "Step #1" on the PhosphOrtholog main page. Input data can be simply copy-pasted/edited/deleted on the user interface (UI) spreadsheet directly in the "Preview for input data set" table. Example input files can also be downloaded through the "download" links immediately below

mapped without PhosphOrtholog, i.e., percentage of novel sites mapped in input data; and (c) the recovery of known orthologous phosphosites (sites annotated in PhosphoSitePlus for that dataset) by PhosphOrtholog. As the system maps the cross-species data, the progress can be assessed through a progress bar above the output columns shown in Step #3 of Fig. 2b, which refreshes every 5 s to reflect the mapping progress.

The five columns in our output file consist of PTM site information of species 1 (first two columns) followed by the PTM information of species 2 (next two columns) and E-value confidence score (last) column. The E-value score is provided if the cross-species site mapped is not reported in PhosphoSitePlus. This E-value represents the probability of sequence alignment occurring by chance. If the mapping was previously known, it returns a message “From PhosphoSitePlus” indicating that it was retrieved from the PhosphoSitePlus database.

Using the example human-mouse data as input, users will find that ~14 % of all mapped sites are novel orthologous sites (85/633). Since data from mouse and human are extensively curated within the PSP database, PhosphOrtholog could directly retrieve the known orthologous sites from PSP. In conclusion, the key feature that differentiates PhosphOrtholog from other mapping tools is its ability to map between hundreds of newly identified PTM sites in different species identified by MS-based proteomics studies, enabling mapping of novel sites of any PTM type through the click of a button.

4 Notes

1. If data from only one species is used as input, PhosphOrtholog cannot continue mapping since prediction is not one of its functions and will return an error asking for the correct input format.
2. One should filter out nonunique records within the input file to PhosphOrtholog.
3. No blank lines are allowed within input file.
4. Records with <90 % localization probability should be filtered out to avoid errors (caused due to “NANA”) from the human example dataset.
5. Special care should be taken to ensure that there are no trailing characters or carriage returns at the end of each line in the input file for PhosphOrtholog. ^M might get added depending on which operating system or software is used to generate the input file. One can check in the terminal shell if using a MAC OSX or Linux. Using vi editor in the terminal, open the file using “vi filename” and then get rid of ^M using the command “:%s/^V^M//g”. The ^v is a CONTROL-V character and ^m is a CONTROL-M.

References

1. Wang Y-C, Peterson SE, Loring JF (2014) Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res* 24:143–160
2. Boersema PJ, Foong LY, Ding VMY et al (2010) In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. *Mol Cell Proteomics* 9:84–99. doi:[10.1074/mcp.M900291-MCP200](https://doi.org/10.1074/mcp.M900291-MCP200)
3. Hornbeck PV, Kornhauser JM, Tkachev S et al (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40:D261–D270. doi:[10.1093/nar/gkr1122](https://doi.org/10.1093/nar/gkr1122)
4. Hornbeck PV, Zhang B, Murray B et al (2014) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res*. doi:[10.1093/nar/gku1267](https://doi.org/10.1093/nar/gku1267)
5. Hornbeck PV, Chabra I, Kornhauser JM et al (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 4:1551–1561. doi:[10.1002/pmic.200300772](https://doi.org/10.1002/pmic.200300772)
6. Landry CR, Levy ED, Michnick SW (2009) Weak functional constraints on phosphoproteomes. *Trends Genet* 25:193–197. doi:[10.1016/j.tig.2009.03.003](https://doi.org/10.1016/j.tig.2009.03.003)
7. Tan CS, Bodenmiller B, Pasulescu A, Jovanovic M, Hengartner MO, Jørgensen C, Bader GD, Aebersold R, Pawson T, Linding R (2009) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal* 2:ra39–ra39
8. Wang Y, Klemke RL (2008) PhosphoBlast, a computational tool for comparing phosphoprotein signatures among large datasets. *Mol Cell Proteomics* 7:145–162. doi:[10.1074/mcp.M700207-MCP200](https://doi.org/10.1074/mcp.M700207-MCP200)
9. Gnad F, Ren S, Cox J et al (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8:R250
10. Gnad F, Gunawardena J, Mann M (2011) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* 39:D253–D260. doi:[10.1093/nar/gkq1159](https://doi.org/10.1093/nar/gkq1159)
11. Trost B, Arsenault R, Griebel P et al (2013) DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites. *Bioinformatics* 29:1693–1695. doi:[10.1093/bioinformatics/btt265](https://doi.org/10.1093/bioinformatics/btt265)
12. Chaudhuri R, Sadrieh A, Hoffman NJ et al (2015) PhosphOrtholog: a web-based tool for cross-species mapping of orthologous protein post-translational modifications. *BMC Genomics* 16:617. doi:[10.1186/s12864-015-1820-x](https://doi.org/10.1186/s12864-015-1820-x)
13. Zhong J, Martinez M, Sengupta S et al (2015) Quantitative phosphoproteomics reveals cross-talk between phosphorylation and O-GlcNAc in the DNA damage response pathway. *Proteomics* 15:591–607. doi:[10.1002/pmic.201400339](https://doi.org/10.1002/pmic.201400339)
14. Hoffman NJ, Parker BL, Chaudhuri R et al (2016) Global phosphoproteomic analysis of human skeletal muscle reveals a network of exercise-regulated kinases and AMPK substrates. *Cell Metab* 22:922–935. doi:[10.1016/j.cmet.2015.09.001](https://doi.org/10.1016/j.cmet.2015.09.001)
15. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech* 26:1367–1372

INDEX

A

Acetylations 23, 25, 168–169, 174, 180,
304, 322, 323, 333, 334, 346–348, 350, 390, 438,
439, 442, 461
Annotation 4, 16, 58, 79, 128, 140, 172, 220, 236,
279, 303, 322, 351, 366, 402, 415, 445, 462
AutoDock Vina 112, 114, 116–120, 122

B

Big Data 4, 11, 27–29, 271
Bioinformatics 3–29, 127–137, 171, 174,
182, 193, 203, 277, 303–318, 321–330, 333–351,
381–392, 399, 403, 408, 417, 424, 432, 439, 460
BioPAX3 274, 275, 278, 282

C

Chemical probes 199
Citryllination 398, 408
Classifications 6, 10, 12, 15, 17, 79–81, 83,
87–91, 94–98, 100, 102, 105, 106, 129, 130, 148,
149, 155, 161–163, 312
Cross-species analysis 348–350
Curation 11, 15, 18, 69, 80, 83, 86, 87, 105,
106, 140, 173, 181, 216, 271, 274–278, 289, 334
Cyberinfrastructure 280, 281, 300
Cysteine 164, 191–196, 198, 200–203,
205, 208–210, 386, 388
Cytoscape 25, 63, 131, 216, 237, 272, 329, 336, 399, 403

D

Data analytics 11, 27, 28
Data integration 11, 15, 27–29, 58, 128, 175, 409
Databases
 protein sequence databases 4, 10, 12–14, 17, 84,
 91, 133, 151, 204, 358, 365–366, 369, 370, 374,
 376–378, 440
 PTM databases 9, 11, 22–25, 154, 171, 173,
 305, 334, 398
 SNP databases 160
 UniProt 4, 41, 68, 84, 130, 140, 164, 194, 215,
 239, 263, 328, 334, 366, 382, 402, 416, 439, 456
Disease associations 9, 160, 173, 176, 181
Domain co-occurrence network 129, 131, 135, 137
Drug discovery 14, 112, 165

F

Function predictions 422, 427, 429, 433
Functional interaction 235–252, 432

G

Glycans 25, 140, 141, 145–154, 156, 165, 166,
174, 179
Glycobioinformatics 139
Glycomics 9, 25, 139–156, 179, 182
Glycoproteomics 149

I

Isoform 12, 41, 57, 177, 239, 324, 336, 366, 415

J

Java 15, 131, 141, 237, 238, 240, 273,
280, 296, 297, 305, 439, 452

K

Kinase 22, 61, 129, 164, 208, 213, 312,
322, 333, 409, 417, 459

M

Mass spectrometry (MS) 13, 52, 128, 141,
173, 195, 236, 304, 322, 350, 357, 381, 395, 437
Molecular docking 112, 113
MS/MS data 401, 407
MS/MS spectra 358, 367, 377, 383, 396,
397, 399, 407
Mutation mapping 9, 256, 263–265
MySQL 14–16, 18, 19, 130, 237, 240–242, 245, 249

N

Naive Bayesian Classifier 278
Network
 isoform network 429, 430, 432
 kinase-substrate network 129, 134, 227
 network-based analysis 237, 245
 phosphatase-substrate network 23, 129,
 131, 134
Non-synonymous single-nucleotide variations
 (nsSNVs) 163, 167, 170, 174–175,
 177–181, 263

O

Open source..... 13, 15, 21, 279
 OpenBabel..... 114, 120, 122, 123
 OpenBEL..... 281, 282, 289
 Open-source..... 113, 141, 236, 245, 276,
 277, 364, 368, 369, 399, 402, 403, 438
 Orthologs 8, 20, 23, 58–60, 129, 306,
 307, 312, 313, 316, 325, 334, 336, 349, 351, 409,
 460–463, 466–468

P

Pan-cancer variomes.....173
 Pathway
 analysis..... 176, 178, 179
 Peptide identifications 22, 194, 367–370,
 372–379, 395, 399, 401
 Phosphatases22–24, 129–135, 165, 177,
 192, 198, 204, 322, 424
 Phosphorylation
 in plants24, 128–130, 132–137, 350
 Post-translational modification (PTM)
 cross-talk314
 site mapping466
 ProSight Lite..... 381–386, 388, 390–392
 Protein..... 111–114, 116, 118–123, 255–260, 262–268
 complexes 14, 57–59, 61, 62, 67, 69,
 74–75, 134, 168, 215, 257, 264, 266, 316,
 322–329, 350, 357, 358
 domains6, 16–18, 79–83, 86–87,
 91, 92, 95, 105, 106, 129, 133, 135, 236,
 242, 329, 409
 folding 80, 419
 function 10, 24, 42, 45, 136, 163, 165,
 177, 213, 305, 307, 309, 312, 333, 402, 406, 417,
 420–422
 identification..... 13, 21, 22, 133, 134, 136,
 194, 198–200, 310, 357–364, 366–379, 395, 402
 interaction..... 9, 11, 21, 129, 131, 134, 177,
 202, 213–217, 219–221, 223–227, 229, 230, 236,
 255–260, 262–268, 275, 276, 299, 300, 305, 306,
 316, 321–330, 334, 402, 421
 interaction network.....129, 131, 276, 300,
 321–330, 402
 mutation 11, 58, 161, 163, 255–260,
 262–268, 329, 351
 network.....129, 131, 134, 136, 255–260, 262–268, 317,
 321–330, 402
 protein family6, 58
 regulation.....313, 316
 sequence..... 4, 41, 61, 79, 133, 151, 162,
 204, 255, 306, 324, 346, 358, 382, 416, 437, 460
 structure
 structure-based virtual screening..... 111–114, 116,
 118–123

 structural matching..... 255–260, 262–268
 structural model prediction.....79
 structural pathway modeling.....255
 Protein ontology 9, 23, 57–64, 67–69, 71–75, 216, 334
 Protein-protein interaction (PPI)
 network..... 129, 131, 134, 136, 255–260,
 262–268, 317, 322, 323, 325, 326, 402
 prediction.....256–260, 262
 Proteoforms 23, 58–59, 61, 62, 66, 68, 69,
 71–74, 216, 334, 336–339, 341, 348, 351, 381,
 388–390, 392, 437
 Proteomics
 top-down 128, 358, 381, 437
 Provenance280, 282, 289–294, 297
 Python..... 273, 280, 297–300

R

Reactive oxygen species (ROS)..... 191–192, 198
 Reactome..... 6, 15–16, 23, 236, 237, 239, 241,
 243–252, 272, 277, 403–405
 ReactomeFIViz 237, 245–247, 251
 Redox state208
 RElational State Transfer Application Programming
 Interface (REST API)280, 296

S

Search engines..... 129, 282, 358, 359, 364–365,
 367–374, 396, 399, 439–441, 444, 454, 455
 Sequence..... 4, 41, 58, 79, 128, 141, 160, 204, 239, 255,
 282, 305, 324, 336, 358, 382, 396, 416, 437, 460
 Sequential window acquisition of all theoretical
 (SWATH)..... 396–398, 403, 405
 S-glutathionylation.....196, 198, 205–207, 209
 Single-nucleotide variations 9, 159–161, 163–182
 Splice isoforms 58, 62, 330, 416, 424, 425,
 430, 431, 434
 Substrate..... 22, 75, 129, 166, 214, 312, 334
 Sulfoxidation191
 Systems biology..... 4, 9, 23, 24, 128–130,
 132–137, 274, 277, 402

T

Text-mining..... 15, 23, 174, 176, 213–217, 219–221,
 223–227, 229, 230, 304, 334, 342, 402, 409

U

UCSF chimera 113–119, 122

V

Virtual screening 111, 113

W

Web application, 280, 466–467