



生命科学前沿及应用生物技术

比较蛋白质组学 (影印版) 的生物信息学

(美) C. H. Wu C. Chen 编著

BIOINFORMATICS FOR
COMPARATIVE PROTEOMICS



科学出版社

生命科学前沿及应用生物技术大系·典藏版

比较蛋白质组学的生物信息学（影印版）

Bioinformatics for Comparative Proteomics

〔美〕C. H. Wu C. Chen 编著

科学出版社

北 京

图字：01-2013-1374 号

内 容 简 介

应用生物技术大系和现代生命科学前沿系列图书分别被列为“十一五”和“十二五”国家重点图书出版规划项目。本丛书针对生命科学领域前沿重点发展方向以及应用生物技术领域的新成果、新思路、新方法和新技术，全面展示了其最新的发展动态，涵盖了基础理论和主要技术方法，呈现了新的概念与理论、技术，在更深层次上阐明了生命的本质规律，给人们提供了新的认识生命本质的手段，也为生物技术服务于人类开辟了新的途径。涉及领域包括生物医药、干细胞技术、工业微生物学、蛋白质及蛋白质组、系统生物学、合成生物学、生物材料、农业生物技术、环境生物技术、海洋生物技术、生物资源与安全等。

Reprint from English language edition: *Bioinformatics for Comparative Proteomics*

By Cathy H. Wu and Chuming Chen

Copyright © 2011 Humana Press

Humana Press is a part of Springer Science+Business Media. All Rights Reserved

This reprint has been authorized by Springer Science & Business Media for distribution in China Mainland only and not for export therefrom.

图书在版编目(CIP)数据

生命科学前沿及应用生物技术大系：典藏版/舒红兵等编著. —北京：科学出版社，2016

ISBN 978-7-03-047487-2

I. ①现… II. ①舒… III. ①生命科学—研究②生物工程—研究 IV. ①Q1-0②Q819

中国版本图书馆 CIP 数据核字(2016)第 043876 号

责任编辑：王 静 李 悦

责任印制：张 伟 / 封面设计：刘新新

科 学 出 版 社 出 版

北京东黄城根北街 16 号

邮政编码：100717

<http://www.sciencep.com>

北京厚诚则铭印刷科技有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 7 月第 一 版 开本：787×1092 1/16

2016 年 7 月第一次印刷 印张：2108

字数：49 985 000

定价：8900.00 元

《新生物学丛书》专家委员会成员名单

主 任：蒲慕明

副主任：吴家睿

专家委员会成员（按姓氏汉语拼音排序）

昌增益	陈洛南	陈晔光	邓兴旺	高 福
韩忠朝	贺福初	黄大昉	蒋华良	金 力
李家洋	林其谁	马克平	孟安明	裴 钢
饶 毅	饶子和	施一公	舒红兵	王 琛
王梅祥	王小宁	吴仲义	徐安龙	许智宏
薛红卫	詹启敏	张先恩	赵国屏	赵立平
钟 扬	周忠和	朱 祯		

《新生物学丛书》丛书序

当前，一场新的生物学革命正在展开。为此，美国国家科学院研究理事会于 2009 年发布了一份战略研究报告，提出一个“新生物学”（New Biology）时代即将来临。这个“新生物学”，一方面是生物学内部各种分支学科的重组与融合，另一方面是化学、物理、信息科学、材料科学等众多非生命学科与生物学的紧密交叉与整合。

在这样一个全球生命科学发展变革的时代，我国的生命科学研究也正在高速发展，并进入了一个充满机遇和挑战的黄金期。在这个时期，将会产生许多具有影响力、推动力的科研成果。因此，有必要通过系统性集成和出版相关主题的国内外优秀图书，为后人留下一笔宝贵的“新生物学”时代精神财富。

科学出版社联合国内一批有志于推进生命科学发展的专家与学者，联合打造了一个 21 世纪中国生命科学的传播平台——《新生物学丛书》。希望通过这套丛书的出版，记录生命科学的进步，传递对生物技术发展的梦想。

《新生物学丛书》下设三个子系列：科学风向标，着重收集科学发展战略和态势分析报告，为科学管理者和科研人员展示科学的最新动向；科学百家园，重点收录国内外专家与学者的科研专著，为专业工作者提供新思想和新方法；科学新视窗，主要发表高级科普著作，为不同领域的研究人员和科学爱好者普及生命科学的前沿知识。

如果说科学出版社是一个“支点”，这套丛书就像一根“杠杆”，那么读者就能够借助这根“杠杆”成为撬动“地球”的人。编委会相信，不同类型的读者都能够从这套丛书中得到新的知识信息，获得思考与启迪。

《新生物学丛书》专家委员会

主 任：蒲慕明

副主任：吴家睿

2012年3月

Preface

With the rapid development of proteomic technologies in life sciences and in clinical applications, many bioinformatics methodologies, databases, and software tools have been developed to support comparative proteomics study. This volume aims to highlight the current status, challenges, open problems, and future trends in developing bioinformatics tools and resources for comparative proteomics research and to serve as a definitive source of reference providing both the breadth and depth needed on the subject of *Bioinformatics for Comparative Proteomics*.

The volume is structured to introduce three major areas of research methods: (1) basic bioinformatics frameworks related to comparative proteomics, (2) bioinformatics databases and tools for proteomics data analysis, and (3) integrated bioinformatics systems and approaches for studying comparative proteomics in the systems biology context.

Part I (Bioinformatics Framework for Comparative Proteomics) consists of seven chapters:

Chapter 1 presents a comprehensive review (with categorization and description) of major protein bioinformatics databases and resources that are relevant to comparative proteomics research.

Chapter 2 provides a practical guide to the comparative proteomics community for exploiting the knowledge captured from and the services provided in UniProt databases.

Chapter 3 introduces the InterPro protein classification system for automatic protein annotation and reviews the signature methods used in the InterPro database.

Chapter 4 introduces the Reactome Knowledgebase that provides an integrated view of the molecular details of human biological processes.

Chapter 5 introduces eFIP (extraction of Functional Impact of Phosphorylation), a Web-based text mining system that can aid scientists in quickly finding abstracts from literature related to the phosphorylation (including site and kinase), interactions, and functional aspects of a given protein.

Chapter 6 presents a tutorial for the Protein Ontology (PRO) Web resources to help researchers in their proteomic studies by providing key information about protein diversity in terms of evolutionary-related protein classes based on full-length sequence conservation and the various protein forms that arise from a gene along with the specific functional annotation.

Chapter 7 describes a method for the annotation of functional residues within experimentally uncharacterized proteins using position-specific site annotation rules derived from structural and experimental information.

Part II (Proteomic Bioinformatics) consists of ten chapters:

Chapter 8 describes how the detailed understanding of information value of mass spectrometry-based proteomics data can be elucidated by performing simulations using synthetic data.

Chapter 9 describes the concepts, prerequisites, and methods required to analyze a shotgun proteomics data set using a tandem mass spectrometry search engine.

Chapter 10 presents computational methods for quantification and comparison of peptides by label-free LC-MS analysis, including data preprocessing, multivariate statistical methods, and detection of differential protein expression.

Chapter 11 proposes an alternative to MS/MS spectrum identification by combining the uninterpreted MS/MS spectra from overlapping peptides and then determining the consensus identifications for sets of aligned MS/MS spectra.

Chapter 12 describes the Trans-Proteomic Pipeline, a freely available open-source software suite that provides uniform analysis of LC-MS/MS data from raw data to quantified sample proteins.

Chapter 13 provides an overview of a set of open-source software tools and steps involved in ELISA microarray data analysis.

Chapter 14 presents the state of the art on the Proteomics Databases and Repositories.

Chapter 15 is a brief guide to preparing both large- and small-scale protein interaction data for publication.

Chapter 16 demonstrates a new graphical user interface tool called PRIDE Converter, which greatly simplifies the submission of MS data to PRIDE database for submitted proteomics manuscripts.

Chapter 17 presents a method for describing a protein's posttranslational modifications by integrating the top-down and bottom-up MS data using the Protein Inference Engine.

Chapter 18 describes an integrated top-down and bottom-up approach facilitated by concurrent liquid chromatography-mass spectrometry analysis and fraction collection for comprehensive high-throughput intact protein profiling.

Part III (Comparative Proteomics in Systems Biology) consists of four chapters:

Chapter 19 gives an overview of the content and usage of the PhosphoPep database, which supports systems biology signaling research by providing interactive interrogation of MS-derived phosphorylation data from four different organisms.

Chapter 20 describes "omics" data integration to map a list of identified proteins to a common representation of the protein and uses the related structural, functional, genetic, and disease information for functional categorization and pathway mapping.

Chapter 21 describes a knowledge-based approach relying on existing metabolic pathway information and a direct data-driven approach for a metabolic pathway-centric integration of proteomics and metabolomics data.

Chapter 22 provides a detailed description of a method used to study temporal changes in the endoplasmic reticulum (ER) proteome of fibroblast cells exposed to ER stress agents (tunicamycin and thapsigargin).

This volume targets the readers who wish to learn about state-of-the-art bioinformatics databases and tools, novel computational methods and future trends in proteomics data analysis, and comparative proteomics in systems biology. The audience may range from graduate students embarking upon a research project, to practicing biologists working on proteomics and systems biology research, and to bioinformaticians developing advanced databases, analysis tools, and integrative systems. With its interdisciplinary nature, this volume is expected to find a broad audience in biotechnology and pharmaceutical companies 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 all the authors and coauthors who had contributed to this volume. We thank our series editor, Dr. John M. Walker, for reviewing all the chapter manuscripts and providing constructive comments. We also thank Dr. Winona C. Barker from Georgetown University for reviewing the manuscripts. We thank Dr. Qinghua Wu for proof reading the book draft. Finally, we would like to extend our thanks to David C. Casey and Anne Meagher of Springer US, Jeya Ruby and Ravi Amina of SPi for their help in the compilation of this book.

Newark, DE, USA

Cathy H. Wu and Chuming Chen

Contributors

- RUEDI AEBERSOLD • *Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland*
- ROLF APWEILER • *The European Bioinformatics Institute, Cambridge, UK*
- CECILIA N. ARIGHI • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- NUNO BANDEIRA • *Center for Computational Mass Spectrometry, University of California, San Diego, La Jolla, CA, USA*
- HARALD BARSNES • *Department of Informatics, University of Bergen, Bergen, Norway*
- BERND BODENMILLER • *Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland*
- KRISTY J. BROWN • *Center for Genetic Medicine Research, Children's National Medical Center, Washington, DC, USA*
- CHUMING CHEN • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- NILS CHRISTIAN • *Max-Planck-Institute for Molecular Plant Physiology, Potsdam-Golm, Germany*
- DON S. DALY • *Pacific Northwest National Laboratory, Richland, WA, USA*
- PETER D'EUSTACHIO • *Department of Biochemistry, New York University School of Medicine, New York, NY, USA*
- OLIVER EBENHÖH • *Max-Planck-Institute for Molecular Plant Physiology, Potsdam-Golm, Germany*
- NATHAN J. EDWARDS • *Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- INGVAR EIDHAMMER • *Department of Informatics, University of Bergen, Bergen, Norway*
- JAN ERIKSSON • *Swedish University of Agricultural Sciences, Uppsala, Sweden*
- DAVID FENYÖ • *The Rockefeller University, New York, NY, USA*
- CATHERINE A. FORMOLO • *Center for Genetic Medicine Research, Children's National Medical Center, Washington, DC, USA*
- MORGAN C. GIDDINGS • *Departments of Microbiology & Immunology and Biomedical Engineering, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
- BRIAN HALLIGAN • *Bioinformatics, Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, USA*
- YETRIB HATHOUT • *Center for Genetic Medicine Research, Children's National Medical Center, Washington, DC, USA*
- HENNING HERMIJAKOB • *EMBL Outstation, European Bioinformatics Institute (EBI), Cambridge, UK*
- HONGZHAN HUANG • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*

- SARAH HUNTER • *EMBL Outstation, European Bioinformatics Institute (EBI), Cambridge, UK*
- STUART R. JEFFERYS • *Department of Bioinformatics & Computational Biology, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
- ROBEL Y. KAHSAY • *DuPont Central Research & Development, Wilmington, DE, USA*
- ANDREW KELLER • *Rosetta Biosoftware, Seattle, WA, USA*
- LENNART MARTENS • *EMBL Outstation, European Bioinformatics Institute (EBI), Cambridge, UK*
- PATRICK MAY • *Max-Planck-Institute for Molecular Plant Physiology, Potsdam-Golm, Germany*
- JENNIFER MCDOWALL • *EMBL Outstation, European Bioinformatics Institute (EBI), Cambridge, UK*
- PETER B. MCGARVEY • *Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- MICHELLE MINTZ • *Center for Genetic Medicine Research, Children's National Medical Center, Washington, DC, USA*
- DARREN A. NATALE • *Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- JULES A. NCHOUTMBOUBE • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- CLAIRE O'DONOVAN • *The European Bioinformatics Institute, Cambridge, UK*
- SANDRA ORCHARD • *EMBL Outstation, European Bioinformatics Institute (EBI), Cambridge, UK*
- LJILJANA PAŠA-TOLIĆ • *Pacific Northwest National Laboratory, Richland, WA, USA*
- FLORIAN REISINGER • *EMBL Outstation, European Bioinformatics Institute (EBI), Cambridge, UK*
- HABTOM W. RESSOM • *Department of Oncology, Georgetown University Medical Center, Washington, DC, USA*
- ERROL W. ROBINSON • *Pacific Northwest National Laboratory, Richland, WA, USA*
- VIJAY K. SHANKER • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- DAVID SHTEYNBERG • *Institute for Systems Biology, Seattle, WA, USA*
- AMY Y. SIU • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- ASAKO TAKANOHASHI • *Center for Genetic Medicine Research, Children's National Medical Center, Washington, DC, USA*
- ZHIXIN TIAN • *Pacific Northwest National Laboratory, Richland, WA, USA*
- NIKOLA TOLIĆ • *Pacific Northwest National Laboratory, Richland, WA, USA*
- CATALINA O. TUDOR • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- ADELINE VANDERVER • *Center for Genetic Medicine Research, Children's National Medical Center, Washington, DC, USA*
- RENCY S. VARGHESE • *Department of Oncology, Georgetown University Medical Center, Washington, DC, USA*

- SONA VASUDEVAN • *Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- C.R. VINAYAKA • *Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*
- JUAN ANTONIO VIZCAÍNO • *EMBL Outstation, European Bioinformatics Institute (EBI), Cambridge, UK*
- DIRK WALTHER • *Max-Planck-Institute for Molecular Plant Physiology, Potsdam-Golm, Germany*
- WOLFRAM WECKWERTH • *Molecular Systems Biology, University of Vienna, Vienna, Austria*
- AMANDA M. WHITE • *Pacific Northwest National Laboratory, Richland, WA, USA*
- CATHY H. WU • *Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA*
- SI WU • *Pacific Northwest National Laboratory, Richland, WA, USA*
- RICHARD C. ZANGAR • *Pacific Northwest National Laboratory, Richland, WA, USA*
- JIAN ZHANG • *Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA*

Contents

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

PART I: BIOINFORMATICS FRAMEWORK FOR COMPARATIVE PROTEOMICS

1 Protein Bioinformatics Databases and Resources	3
<i>Chuming Chen, Hongzhan Huang, and Cathy H. Wu</i>	
2 A Guide to UniProt for Protein Scientists	25
<i>Claire O'Donovan and Rolf Apweiler</i>	
3 InterPro Protein Classification	37
<i>Jennifer McDowall and Sarah Hunter</i>	
4 Reactome Knowledgebase of Human Biological Pathways and Processes.	49
<i>Peter D'Eustachio</i>	
5 eFIP: A Tool for Mining Functional Impact of Phosphorylation from Literature.	63
<i>Cecilia N. Arighi, Amy Y. Siu, Catalina O. Tudor, Jules A. Nchoutmboube, Cathy H. Wu, and Vijay K. Shanker</i>	
6 A Tutorial on Protein Ontology Resources for Proteomic Studies	77
<i>Cecilia N. Arighi</i>	
7 Structure-Guided Rule-Based Annotation of Protein Functional Sites in UniProt Knowledgebase	91
<i>Sona Vasudevan, C.R. Vinayaka, Darren A. Natale, Hongzhan Huang, Robel Y. Kahsay, and Cathy H. Wu</i>	

PART II: PROTEOMIC BIOINFORMATICS

8 Modeling Mass Spectrometry-Based Protein Analysis	109
<i>Jan Eriksson and David Fenyő</i>	
9 Protein Identification from Tandem Mass Spectra by Database Searching	119
<i>Nathan J. Edwards</i>	
10 LC-MS Data Analysis for Differential Protein Expression Detection	139
<i>Rency S. Varghese and Habtom W. Ressom</i>	
11 Protein Identification by Spectral Networks Analysis.	151
<i>Nuno Bandeira</i>	
12 Software Pipeline and Data Analysis for MS/MS Proteomics: The Trans-Proteomic Pipeline.	169
<i>Andrew Keller and David Shteynberg</i>	

13	Analysis of High-Throughput ELISA Microarray Data	191
	<i>Amanda M. White, Don S. Daly, and Richard C. Zangar</i>	
14	Proteomics Databases and Repositories.	213
	<i>Lennart Martens</i>	
15	Preparing Molecular Interaction Data for Publication	229
	<i>Sandra Orchard and Henning Hermjakob</i>	
16	Submitting Proteomics Data to PRIDE Using PRIDE Converter.	237
	<i>Harald Barsnes, Juan Antonio Vizcaino, Florian Reisinger, Ingvar Eidhammer, and Lennart Martens</i>	
17	Automated Data Integration and Determination of Posttranslational Modifications with the Protein Inference Engine	255
	<i>Stuart R. Jefferys and Morgan C. Giddings</i>	
18	An Integrated Top-Down and Bottom-Up Strategy for Characterization of Protein Isoforms and Modifications.	293
	<i>Si Wu, Nikola Tolić, Zhixin Tian, Errol W. Robinson, and Ljiljana Paša-Tolić</i>	
PART III: COMPARATIVE PROTEOMICS IN SYSTEMS BIOLOGY		
19	Phosphoproteome Resource for Systems Biology Research	307
	<i>Bernd Bodenmiller and Ruedi Aebersold</i>	
20	Protein-Centric Data Integration for Functional Analysis of Comparative Proteomics Data.	323
	<i>Peter B. McGarvey, Jian Zhang, Darren A. Natale, Cathy H. Wu, and Hongzhan Huang</i>	
21	Integration of Proteomic and Metabolomic Profiling as well as Metabolic Modeling for the Functional Analysis of Metabolic Networks	341
	<i>Patrick May, Nils Christian, Oliver Ebenhöf, Wolfram Weckwerth, and Dirk Walther</i>	
22	Time Series Proteome Profiling	365
	<i>Catherine A. Formolo, Michelle Mintz, Asako Takanohashi, Kristy J. Brown, Adeline Vanderver, Brian Halligan, and Yetrib Hathout</i>	
	<i>Index</i>	379

Part I

Bioinformatics Framework for Comparative Proteomics

Chapter 1

Protein Bioinformatics Databases and Resources

Chuming Chen, Hongzhan Huang, and Cathy H. Wu

Abstract

In the past decades, a variety of publicly available data repositories and resources have been developed to support protein related information management, data-driven hypothesis generation and biological knowledge discovery. However, there is also an increasing confusion for the researchers who are trying to quickly find the appropriate resources to help them solve their problems. In this chapter, we present a comprehensive review (with categorization and description) of major protein bioinformatics databases and resources that are relevant to comparative proteomics research. We conclude the chapter by discussing the challenges and opportunities for developing new protein bioinformatics databases.

Key words: Bioinformatics, Database, Protein sequence, Protein family, Protein structure, Protein function, Proteomics, Data integration, Comparative analysis

1. Introduction

Advances of high-throughput technologies in the study of molecular biology systems in the past decades have marked the beginning of a new era of research, in which biological researchers systematically study organisms on the levels of genomes (complete genetic sequences) (1), transcriptomes (gene expressions) (2) and proteomes (protein expressions) (3). Because proteins occupy a middle ground molecularly between gene and transcript information and higher levels of molecular and cellular structure and organization, and most physiological and pathological processes are manifested at the protein level, biological scientists are growingly interested in applying proteomics techniques to foster a better understanding of basic molecular biology, disease processes and discovery of new diagnostic, prognostic and therapeutic targets for numerous diseases (4, 5).

Recently, proteomics data analysis has moved toward information integration of multiple studies including cross-species analyses (6–9). The richness of proteomics data allows researchers to ask complex biological questions and gain new scientific insights. To support comparative proteomics, data-driven hypothesis generation, and biological knowledge discovery, many protein-related bioinformatics databases, query facilities, and data analysis software tools have been developed. These organize and provide biological annotations for individual proteins to support sequence, structural, functional and evolutionary analyses in the context of pathway, network and systems biology. However, it is not always easy for researchers to quickly find the pieces of related information. In this chapter, we present a comprehensive review (with categorization and description) of major protein bioinformatics databases and resources that are relevant to comparative proteomics research. We highlight some of these databases, and focus on the types of data stored and related data access and data analysis supports. We also discuss the challenges and opportunities for developing new protein bioinformatics databases in terms of supporting data integration and comparative analysis, maintaining data provenance and managing biological knowledge.

2. Overview

Our coverage of protein bioinformatics databases in this chapter is by no means exhaustive. We refer the readers to ref. 10 for a more complete list. Our intention is to cover those that are recent, high quality, publicly available, and are expected to be of interest to more users in the comparative proteomics community. Based on the topics and data stored, protein bioinformatics databases can be primarily classified as sequence databases, family databases, structure databases, function databases and proteomics databases as shown in Table 1. It is worth noting that certain databases can be classified into more than one category. Please visit <http://www.proteininformationresource.org/staff/chenc/MiMB/dbSummary.html> to access the databases reviewed in this chapter through their corresponding web addresses (URLs).

3. Databases and Resources Highlights

3.1. Protein Sequence Databases

Protein sequence databases serve as the archival repositories for collections of protein sequences as well as their associated annotations. These databases are also the primary sources for developing other