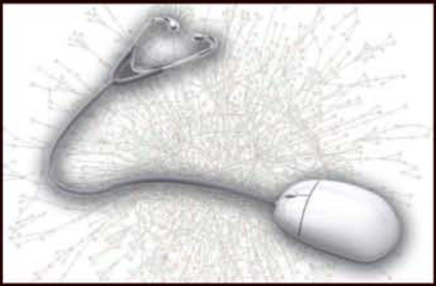
Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Biological Data Mining

CEAL-1 19q13.1 carcinoembryonic antigen
BCL3 19q13.1 B-cell lymphoma
CBLC 19q13.2 Ectopic CAS
LU 19q13.2 Lutheran blood groop antig
PVRL2 19q13.2 poliovirus receptor
TOMM 19q13.2 translocase
APO E 19q13.2 apolipoprotein E
APO CI 19q13.2
APO C4 19q13.2
APO C2 19q13.2
CLPTM 19q13.2
GEMIN 19q13.3
HPP1 19q13.3
Alzheimer



Edited by
Jake Y. Chen
Stefano Lonardi

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Biological Data Mining

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

SERIES EDITOR

Vipin Kumar

University of Minnesota
Department of Computer Science and Engineering
Minneapolis, Minnesota, U.S.A

AIMS AND SCOPE

This series aims to capture new developments and applications in data mining and knowledge discovery, while summarizing the computational tools and techniques useful in data analysis. This series encourages the integration of mathematical, statistical, and computational methods and techniques through the publication of a broad range of textbooks, reference works, and handbooks. The inclusion of concrete examples and applications is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of data mining and knowledge discovery methods and applications, modeling, algorithms, theory and foundations, data and knowledge visualization, data mining systems and tools, and privacy and security issues.

PUBLISHED TITLES

UNDERSTANDING COMPLEX DATASETS: Data Mining with Matrix Decompositions
David Skillicorn

COMPUTATIONAL METHODS OF FEATURE SELECTION
Huan Liu and Hiroshi Motoda

CONSTRAINED CLUSTERING: Advances in Algorithms, Theory, and Applications
Sugato Basu, Ian Davidson, and Kiri L. Wagstaff

KNOWLEDGE DISCOVERY FOR COUNTERTERRORISM AND LAW ENFORCEMENT
David Skillicorn

MULTIMEDIA DATA MINING: A Systematic Introduction to Concepts and Theory
Zhongfei Zhang and Ruofei Zhang

NEXT GENERATION OF DATA MINING
Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar

DATA MINING FOR DESIGN AND MARKETING
Yukio Ohsawa and Katsutoshi Yada

THE TOP TEN ALGORITHMS IN DATA MINING
Xindong Wu and Vipin Kumar

GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY, Second Edition
Harvey J. Miller and Jiawei Han

TEXT MINING: CLASSIFICATION, CLUSTERING, AND APPLICATIONS
Ashok N. Srivastava and Mehran Sahami

BIOLOGICAL DATA MINING
Jake Y. Chen and Stefano Lonardi

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Biological Data Mining

Edited by
Jake Y. Chen
Stefano Lonardi

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-8684-3 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Biological data mining / editors, Jake Y. Chen, Stefano Lonardi.
p. cm. -- (Data mining and knowledge discovery series)
Includes bibliographical references and index.
ISBN 978-1-4200-8684-3 (hardcover : alk. paper)
1. Bioinformatics. 2. Data mining. 3. Computational biology. I. Chen, Jake. II. Lonardi, Stefano. III. Title. IV. Series.

QH324.2.B578 2010
570.285--dc22 2009028067

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

Modern biology has become an information science. Since the invention of a DNA sequencing method by Sanger in the late seventies, public repositories of genomic sequences have been growing exponentially, doubling in size every 16 months—a rate often compared to the growth of semiconductor transistor densities in CPUs known as Moore's Law. In the nineties, the public–private race to sequence the human genome further intensified the fervor to generate high-throughput biomolecular data from highly parallel and miniaturized instruments. Today, sequencing data from thousands of genomes, including plants, mammals, and microbial genomes, are accumulating at an unprecedented rate. The advent of second-generation DNA sequencing instruments, high-density cDNA microarrays, tandem mass spectrometers, and high-power NMRs have fueled the growth of molecular biology into a wide spectrum of disciplines such as personalized genomics, functional genomics, proteomics, metabolomics, and structural genomics. Few experiments in molecular biology and genetics performed today can afford to ignore the vast amount of biological information publicly accessible. Suddenly, molecular biology and genetics have become data rich.

Biological data mining is a *data-guzzling turbo engine* for postgenomic biology, driving the competitive race toward unprecedented biological discovery opportunities in the twenty-first century. Classical bioinformatics emerged from the study of macromolecules in molecular biology, biochemistry, and biophysics. Analysis, comparison, and classification of DNA and protein sequences were the dominant themes of bioinformatics in the early nineties. Machine learning mainly focused on predicting genes and proteins functions from their sequences and structures. The understanding of cellular functions and processes underlying complex diseases were out of reach. Bioinformatics scientists were a rare breed, and their contribution to molecular biology and genetics was considered marginal, because the computational tools available then for biomolecular data analysis were far more primitive than the array of experimental techniques and assays that were available to life scientists. Today, we are now witnessing the reversal of these past trends. Diverse sets of data types that cover a broad spectrum of genotypes and phenotypes, particularly those related to human health and diseases, have become available. Many interdisciplinary researchers, including applied computer scientists, applied mathematicians, biostatisticians, biomedical researchers, clinical scientists, and biopharmaceutical professionals, have discovered in biology a *gold* *mine* of knowledge leading to many exciting possibilities: the unraveling of the tree of life, harnessing the power of microbial organisms for renewable energy, finding new ways to diagnose disease early, and developing new therapeutic compounds that save lives. Much of the experimental high-throughput biology data are generated and analyzed "in haste," therefore leaving plenty of opportunities for knowledge discovery even after the original data are released. Most of the bets on the race to *separate the wheat from the chaff* have been placed on biological data mining techniques. After all, when easy, straightforward, first-pass data analysis has not yielded novel biological insights, data mining techniques must be able to help—or, many presumed so.

In reality, biological data mining is still much of an "art," successfully practiced by a few bioinformatics research groups that occupy themselves with solving real-world biological problems. Unlikely data mining in business, where the major concerns are often related to the bottom line—profit—the goals of biological data mining can be as diverse as the spectrum of biological questions that exist. In the business domain, association rules discovered between sales items are immediately actionable; in biology, any unorthodox hypothesis produced by computational models has to be first red-flagged and is lucky to be validated experimentally. In the Internet business domain, classification, clustering, and visualization of blogs, network traffic patterns, and news feeds add significant values to regular Internet users who are unaware of high-level patterns that may exist in the data set; in molecular biology and genetics, any clustering or classification of the data presented to biologists may promptly elicit questions like "great, but how and why did it happen?" or "how can you explain these results in the context of the biology I know?" The majority of general-purpose data mining techniques do not take into consideration the prior knowledge domain of the biological problem, leading them to often underperform hypothesis-driven biological investigative techniques. The high level of variability of measurements inherent in many types of biological experiments or samples, the general unavailability of experimental replicates, the large number of hidden variables in the data, and the high correlation of biomolecular expression measurements also constitute significant challenges in the application of classical data mining methods in biology. Many biological data mining projects are attempted and then abandoned, even by experienced data mining scientists. In the extreme cases, large-scale biological data mining efforts are jokingly labeled as *fishing expeditions* and dispelled, in national grant proposal review panels.

This book represents a culmination of our past research efforts in biological data mining. Throughout this book, we wanted to showcase a small, but noteworthy sample of successful projects involving data mining and molecular biology. Each chapter of the book is authored by a distinguished team of bioinformatics scientists whom we invited to offer the readers the widest possible range of application domains. To ensure high-quality standards, each contributed chapter went through standard peer reviews and a round of revisions. The contributed chapters have been grouped into five major sections.

The first section, entitled *Sequence, Structure, and Function*, collects contributions on data mining techniques designed to analyze biological sequences and structures with the objective of discovering novel functional knowledge. The second section, on *Genomics, Transcriptomics, and Proteomics*, contains studies addressing emerging large-scale data mining challenges in analyzing high-throughput "omics" data. The chapters in the third section, entitled *Functional and Molecular Interaction Networks*, address emerging system-scale molecular properties and their relevance to cellular functions. The fourth section is about *Literature, Ontology, and Knowledge Integrations*, and it collects chapters related to knowledge representation, information retrieval, and data integration for structured and unstructured biological data. The contributed works in the fifth and last section, entitled *Genome Medicine Applications*, address emerging biological data mining applications in medicine.

We believe this book can serve as a valuable guide to the field for graduate students, researchers, and practitioners. We hope that the wide range of topics covered will allow readers to appreciate the extent of the impact of data mining in molecular biology and genetics. For us, research in data mining and its applications to biology and genetics is fascinating and rewarding. It may even help to save human lives one day. This field offers great opportunities and rewards if one is prepared to learn molecular biology and genetics, design user-friendly software tools under the proper biological assumptions, and validate all discovered hypotheses rigorously using appropriate models.

In closing, we would like to thank all the authors that contributed a chapter in the book. We are also indebted to Randi Cohen, our outstanding publishing editor. Randi efficiently managed timelines and deadlines, gracefully handled the communication with the authors and the reviewers, and took care of every little detail associated with this project. This book could not have been possible without her. Our thanks also go to our families for their support throughout the book project.

Jake Y. Chen
Indianapolis, Indiana

Stefano Lonardi
Riverside, California

Editors

Jake Chen is an assistant professor of informatics at Indiana University School of Informatics and assistant professor of computer science at Purdue School of Science, Indiana. He is the founding director of the Indiana Center for Systems Biology and Personalized Medicine—the first research center in the region to promote the development of systems biology tools towards solving future personalized medicine problems. He is an IEEE senior member and a member of several other interdisciplinary Indiana research centers, including: Center for Computational Biology and Bioinformatics, Center for Bio-computing, Indiana University Cancer Center, and Indiana Center for Environmental Health. He was a scientific co-founder and chief informatics officer (2006–2008) of Predictive Physiology and Medicine, Inc. and the founder of Medeolinx, LLC-Indiana biotech startups developing businesses in emerging personalized medicine and translational bioinformatics markets.

Dr. Chen received PhD and MS degrees in computer science from the University of Minnesota at Twin Cities and a BS in molecular biology and biochemistry from Peking University in China. He has extensive industrial research and management experience (1998–2003), including developing commercial GeneChip microarrays at Affymetrix, Inc. and mapping the first human protein interactome at Myriad Proteomics. After rejoining academia in 2004, he concentrated his research on "translational bioinformatics," studies aiming to bridge the gaps between bioinformatics research and human health applications. He has over 60 publications in the areas of biological data management, biological data mining, network biology, systems biology, and various disease-related omics applications.

Stefano Lonardi is associate professor of computer science and engineering at the University of California, Riverside. He is also a faculty member of the graduate program in genetics, genomics and bioinformatics, the Center for Plant Cell Biology, the Institute for Integrative Genome Biology, and the graduate program in cell, molecular and developmental biology.

Dr. Lonardi received his "Laurea cum laude" from the University of Pisa in 1994 and his PhD, in the summer of 2001, from the Department of Computer Sciences, Purdue University, West Lafayette, IN. He also holds a PhD in electrical and information engineering from the University of Padua (1999). During the summer of 1999, he was an intern at Celera Genomics, Department of Informatics Research, Rockville, MD.

Dr. Lonardi's recent research interests include designing of algorithms, computational molecular biology, data compression, and data mining. He has published more than 30 papers in major theoretical computer science and computational biology journals and has about 45 publications in refereed international conferences. In 2005, he received the CAREER award from the National Science Foundation.

Contributors

Muhammad Abulaish
Department of Computer Science
Jamia Millia Islamia
New Delhi, India

Alberto Apostolico
College of Computing
Georgia Institute of Technology
Atlanta, Georgia

Simon Beaulah
InforSense, Ltd.
London, United Kingdom

Paola Bertolazzi
Istituto di Analisi dei Sistemi ed
Informatica Antonio Ruberti
Consiglio Nazionale delle Ricerche
Rome, Italy

Paul E. Blower
Department of Pharmacology
Ohio State University
Columbus, Ohio

Charles Buck
Bindley Bioscience Center
Purdue University
West Lafayette, Indiana

Jennifer Cai
Department of Pathology
University of Texas Southwestern
Medical Center
Dallas, Texas

Dongsheng Che
Department of Computer Science
East Stroudsburg University
East Stroudsburg, Pennsylvania

Yi-Ping Phoebe Chen
School of Information Technology
Deakin University
Melbourne, Australia

Hyeyoung Cho
Bindley Bioscience Center
Purdue University
West Lafayette, Indiana
and
Department of Bio and Brain
Engineering
KAIST
Daejeon, South Korea

Jeong-Hyeon Choi
Center for Genomics and
Bioinformatics and
School of Informatics
Indiana University
Bloomington, Indiana

Giovanni Ciriello
Department of Information
Engineering
University of Padova
Padova, Italy

Matteo Comin
Department of Information Engineering
University of Padua
Padova, Italy

Mick Correll
InforSense, LLC
Cambridge, Massachusetts

John Crispino
Hematology Oncology
Northwestern University
Chicago, Illinois

Carl Dahlke
Health Information Systems
Northrop Grumman, Inc.
Rockville, Maryland

Mehmet Dalkilic
School of Informatics
Indiana University
Bloomington, Indiana

Lipika Dey
Innovation Labs
Tata Consultancy Services
New Delhi, India

Patrick Dunn
Health Information Systems
Northrop Grumman, Inc.
Rockville, Maryland

Giovanni Felici
Istituto di Analisi dei Sistemi ed Informatica Antonio Ruberti
Consiglio Nazionale delle Ricerche
Rome, Italy

Raffaele Giancarlo
Dipartimento di Matematica ed Applicazioni
University of Palermo
Palermo, Italy

Concettina Guerra
College of Computing
Georgia Institute of Technology
Atlanta, Georgia and
Department of Information Engineering
University of Padua
Padova, Italy

Yike Guo
InforSense, Ltd.
London, United Kingdom

Herb Hagler
Department of Pathology
University of Texas Southwestern Medical Center
Dallas, Texas

Jing-Dong J. Han
Key Laboratory of Molecular Developmental Biology
Center for Molecular Systems Biology
Institute of Genetics and Developmental Biology
Chinese Academy of Sciences
Beijing, People's Republic of China

Christine E. Heitsch
School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia

Hai Hu
Windber Research Institute
Windber, Pennsylvania

Yang Huang
National Institutes of Health
Bethesda, Maryland

Zan Huang
Hematology Oncology
Northwestern University
Chicago, Illinois

Hongmei Jiang
Department of Statistics
Northwestern University
Evanston, Illinois

David Karp
Division of Rheumatology
University of Texas Southwestern
Medical Center
Dallas, Texas

George Karypis
Deparment of Computer Science
University of Minnesota
Minneapolis, Minnesota

Weimao Ke
University of North Carolina
Chapel Hill, North Carolina

Daisuke Kihara
Department of Biological Sciences
and Department of Computer
Science
Markey Center for Structural Biology
College of Science
Purdue University
West Lafayette, Indiana

Sun Kim
Center for Genomics and
Bioinformatics and School of
Informatics
Indiana University
Bloomington, Indiana

Megan Kong
Department of Pathology
University of Texas Southwestern
Medical Center
Dallas, Texas

Yazhene Krishnaraj
Wayne State University
Detroit, Michigan

Giuseppe Lancia
Dipartimento di Matematica e
Informatica
University of Udine
Udine, Italy

Chia-Ju Lee
Biomedical Informatics Center
Northwestern University
Chicago, Illinois

Jamie Lee
Department of Pathology
University of Texas Southwestern
Medical Center
Dallas, Texas

Gang Li
School of Information Technology
Deakin University
Melbourne, Australia

Guojun Li
Department of Biochemistry and Molecular Biology and Institute of Bioinformatics
University of Georgia
Athens, Georgia
and
School of Mathematics and System Sciences
Shandong University
Jinan, People's Republic of China

Li Liao
Computer and Information Sciences
University of Delaware
Newark, Delaware

Simon Lin
Biomedical Informatics Center
Northwestern University
Chicago, Illinois

Elizabeth McClellan
Division of Biomedical Informatics
University of Texas Southwestern Medical Center
Dallas, Texas
and
Department of Statistical Science
Southern Methodist University
Dallas, Texas

Monnie McGee
Department of Statistical Science
Southern Methodist University
Dallas, Texas

Yehia Mechref
National Center for Glycomics and Glycoproteomics
Department of Chemistry
Indiana University
Bloomington, Indiana

Tijana Milenković
Department of Computer Science
University of California
Irvine, California

Jason H. Moore
Computational Genetics Laboratory
Norris-Cotton Cancer Center
Departments of Genetics and Community and Family Medicine
Dartmouth Medical School
Lebanon, New Hampshire

and
Department of Computer Science
University of New Hampshire
Durham, New Hampshire

and
Department of Computer Science
University of Vermont
Burlington, Vermont

and
Translational Genomics Research Institute
Phoenix, Arizona

Javed Mostafa
University of North Carolina
Chapel Hill, North Carolina

Robin Munro
InforSense, Ltd.
London, United Kingdom

Glenn J. Myatt
Myatt & Johnson, Inc.
Jasper, Georgia

Chris J. Needham
School of Computing
University of Leeds
Leeds, United Kingdom

Cheolhwan Oh
Bindley Bioscience Center
Purdue University
West Lafayette, Indiana

Mihai Pop
Center for Bioinformatics and
Computational Biology
University of Maryland
College Park, Maryland

Teresa M. Przytycka
National Institutes of Health
Bethesda, Maryland

Nataša Pržulj
Department of Computer Science
University of California
Irvine, California

Yu Qian
Department of Pathology
University of Texas Southwestern
Medical Center
Dallas, Texas

Naren Ramakrishnan
Department of Computer Science
Virginia Tech
Blacksburg, Virginia

Huzefa Rangwala
Department of Computer Science
George Mason University
Fairfax, Virginia

Chandan Reddy
Wayne State University
Detroit, Michigan

Catherine P. Riley
Bindley Bioscience Center
Purdue University
West Lafayette, Indiana

Jia Rong
School of Information Technology
Deakin University
Melbourne, Australia

Lee Sael
Department of Computer Science
Purdue University
West Lafayette, Indiana

Davide Scaturro
Dipartimento di Matematica
ed Applicazioni
University of Palermo
Palermo, Italy

Richard H. Scheuermann
Department of Pathology
Division of Biomedical Informatics
University of Texas Southwestern
Medical Center
Dallas, Texas

Kazuhiro Seki
Organization of Advanced Science
and Technology
Kobe University
Kobe, Japan

Jonathan Sheldon
InforSense Ltd.
London, United Kingdom

Barry Smith
Department of Philosophy
University at Buffalo
Buffalo, New York

Junilda Spirollari
New Jersey Institute of Technology
Newark, New Jersey

Burke Squires
Department of Pathology
University of Texas Southwestern Medical Center
Dallas, Texas

Haixu Tang
School of Informatics
National Center for Glycomics and Glycoproteomics
Indiana University
Bloomington, Indiana

Jahiruddin
Department of Computer Science
Jamia Millia Islamia
New Delhi, India

Filippo Utro
Dipartimento di Matematica ed Applicazioni
University of Palermo
Palermo, Italy

Jason T. L. Wang
New Jersey Institute of Technology
Newark, New Jersey

Jeff Wiser
Health Information Systems
Northrop Grumman, Inc.
Rockville, Maryland

Mohammed Zaki
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York

Giuseppe Zanotti
Department of Biological Chemistry
University of Padua
Padova, Italy

Xiang Zhang
Department of Chemistry
Center of Regulatory and Environmental Analytical Metabolomics
University of Louisville
Louisville, Kentucky

Jie Zheng
National Institutes of Health
Bethesda, Maryland

Part I

Sequence, Structure, and Function

Chapter 1

Consensus Structure Prediction for RNA Alignments

Junilda Spirollari and Jason T. L. Wang

New Jersey Institute of Technology

1.1 Introduction

RNA secondary structure prediction has been studied for quite awhile. Many minimum free energy (MFE) methods have been developed for predicting the secondary structures of single RNA sequences, such as mfold [1], RNAfold [2], MPGAfold [3], as well as recent tools presented in the literature [4, 5]. However, the accuracy of predicted structures is far from perfect. As evaluated by Gardner and Giegerich [6], the accuracy of the MFE methods for single sequences is 73% when averaged over many different RNAs.

Recently, a new concept of energy density for predicting the secondary structures of single RNA sequences was introduced [7]. The normalized free energy, or energy density, of an RNA substructure is the free energy of that substructure divided by the length of its underlying sequence. A dynamic programming algorithm, called Densityfold, was developed, which delocalizes the thermodynamic cost of computing RNA substructures and improves on secondary structure prediction via energy density minimization [7]. Here, we extend the concept used in Densityfold and present a tool, called RSpredict, for RNA secondary structure prediction. RSpredict computes the RNA structure with minimum energy density based on the loop decomposition scheme used in the nearest neighbor energy model [8]. RSpredict focuses on the loops in an RNA secondary structure, whereas Densityfold considers RNA substructures where a substructure may contain several loops.

While the energy density model creates a foundation for RNA secondary structure prediction, there are many limitations in Densityfold, just like in all other single sequence-based MFE methods. Optimal structures predicted by these methods do not necessarily represent real structures [9]. This happens due to several reasons. The thermodynamic model may not be accurate. The bases of structural RNAs may be chemically modified and these processes are not included in the prediction model. Finally, some functional RNAs may not have stable secondary structures [6]. Thus, a more reliable approach is to use comparative analysis to compute consensus secondary structures from multiple related RNA sequences [9].

In general, there are three strategies with the comparative approach. The first strategy is to predict the secondary structures of individual RNA sequences separately and then align the structures. Tools such as RNAshapes [10,11], MARNA [12], STRUCTURELAB [13], and RADAR [14,15] are based on this strategy. RNA Sampler [9] and comRNA [16] compare and find stems conserved across multiple sequences and then assemble conserved stem blocks to form consensus structures, in which pseudoknots are allowed.

The second strategy predicts common secondary structures of two or more RNA sequences through simultaneous alignment and consensus structure inference. Tools based on this strategy include RNAscf [17], Foldalign [18], Dynalign [19], stemloc [20], PMcomp [21], MASTR [22], and CARNAC [23]. These tools utilize either folding free energy change parameters or stochastic context-free grammars (SCFGs) and are considered derivations of Sankoff's method [24].

The third strategy is to fold multiple sequence alignments. RNAalifold [25, 26] uses a dynamic programming algorithm to compute the consensus secondary structure with MFE by taking into account thermodynamic stability, sequence covariation together with RIBOSUM-like scoring matrices [27]. Pfold [28] is a SCFG algorithm that produces a prior probability distribution of RNA structures. A maximum likelihood approach is used to estimate a phylogenetic tree for predicting the most likely structure for input sequences. A limitation of Pfold is that it does not run on alignments of more than 40 sequences and in some cases produces no structures due to under-flow errors [6]. Maximum weighted matching (MWM), based on a graph-theoretical approach and developed by Cary and Stormo [29] and Tabaska et al. [30], is able to predict common secondary structures allowing pseudo-knots. KNetFold [31] is a recently published machine learning method, implemented using a hierarchical network of k-nearest neighbor classifiers that analyzes the base pairings of alignment columns in the input sequences through their mutual information, Watson–Crick base pairing rules and thermodynamic base pair propensity derived from RNAfold [2]. The method presented in this chapter, RSpredict, joins the many tools using the third strategy; it accepts a multiple alignment of RNA sequences as input data and predicts the consensus secondary structure for the input sequences via energy density minimization and covariance score calculation.

We also considered two variants of RSpredict, referred to as RSefold and RSdfold respectively. Both RSefold and RSdfold use the same covariance score calculation as in RSpredict. The differences among the three approaches lie in the folding algorithms they adopt. Rse-fold predicts the consensus secondary structure for the input sequences via free energy minimization, as opposed to energy density minimization used in RSpredict. RSdfold does the prediction via energy density minimization, though its energy density is calculated based on RNA substructures as in Densityfold, rather than based on the loops used in RSpredict.

The rest of the chapter is organized as follows. We first describe the implementation and algorithms used by RSpredict, and analyze the time complexity of the algorithms (see Section 1.2). We then present experimental results of running the RSpredict tool as well as comparison with the existing tools (see Section 1.3). The experiments were performed on a variety of datasets. Finally we discuss some properties of RSpredict, possible ways to improve the tool and point out some directions for future research (see Section 1.4).

1.2 Algorithms

RSpredict, which can be freely downloaded from http://datalab.njit.edu/biology/RSpredict, was implemented in the Java programming language. The program accepts, as input data, a multiple sequence alignment in the FASTA or ClustalW format and outputs the consensus secondary structure of the input sequences in both the Vienna style dot bracket format [26] and the connectivity table format [32]. Below, we describe the energy density model adopted by RSpredict. We then present a dynamic programming algorithm for folding a single RNA sequence via energy density minimization. Next, we describe techniques for calculating covariance scores based on the input alignment. Finally we summarize the algorithms used by RSpredict, combining both the folding technique and the covariance scores obtained from the input alignment, and show its time complexity.

1.2.1 Folding of a single RNA sequence

1.2.1.1 Preliminaries

We represent an RNA secondary structure as a fully decomposed set of loops. In general, a loop L can be one of the following (see Figure 1.1):

i. A hairpin loop (which is a loop enclosed by only one base pair; the smallest possible hairpin loop consists of three nucleotides enclosed by a base pair)

ii. A stack, composed of two consecutive base pairs

iii. A bulge loop, if two base pairs are separated only on one side by one or more unpaired bases

iv. An internal loop, if two base pairs are separated by one or more unpaired bases on both sides

v. A multibranched loop, if more than two base pairs are separated by zero or more unpaired bases in the loop

We now introduce some terms and definitions. Let S be an RNA sequence consisting of nucleotides or bases A, U, C, G. $S[i]$ denotes the base at position i of the sequence S and $S[i, j]$ is the subsequence starting at position i and ending at position j in S. A base pair between nucleotides at positions i and j is denoted as (i, j) or $(S[i], S[j])$, and its enclosed sequence is $S[i, j]$. Given a loop L in the secondary structure R of sequence S, the base pair (i^*, j^*) in L is called the *exterior pair* of L if $S[i^*](S[j^*]$, respectively) is closest to the $5'$ ($3'$, respectively) end of R among all nucleotides in L. All other nonexterior base pairs in L are called *interior pairs* of L. The length of a loop L is the number of nucleotides in L. Note that two loops may overlap on a base pair. For example, the interior pair of a stack may be the exterior pair of another stack, or the exterior pair of a hairpin loop. Also note that a bulge or an internal loop has exactly one exterior pair and one interior pair.

We use the energy density concept as follows. Given a secondary structure R, every base pair (i, j) in R is the exterior pair of some loop L. We assign (i, j) and L an energy density, which is the free energy of the loop L divided by the length of L. The set of free energy parameters for nonmultibranched loops used in our algorithm is acquired from [33]. The free energy of a multibranched loop is computed based on the approach adopted by mfold [1], which is a linear function of the number of unpaired bases and the number of base pairs inside the loop, namely $a + b \times n_1 + c \times n_2$, where a, b, c are constants, n_1 is the number of unpaired bases and n_2 is the number of base pairs inside the multibranched loop. We adopt the loop decomposition scheme used in the nearest neighbor energy model developed by Turner et al. [8]. The secondary structure R contains multiple loop components and the energy densities of

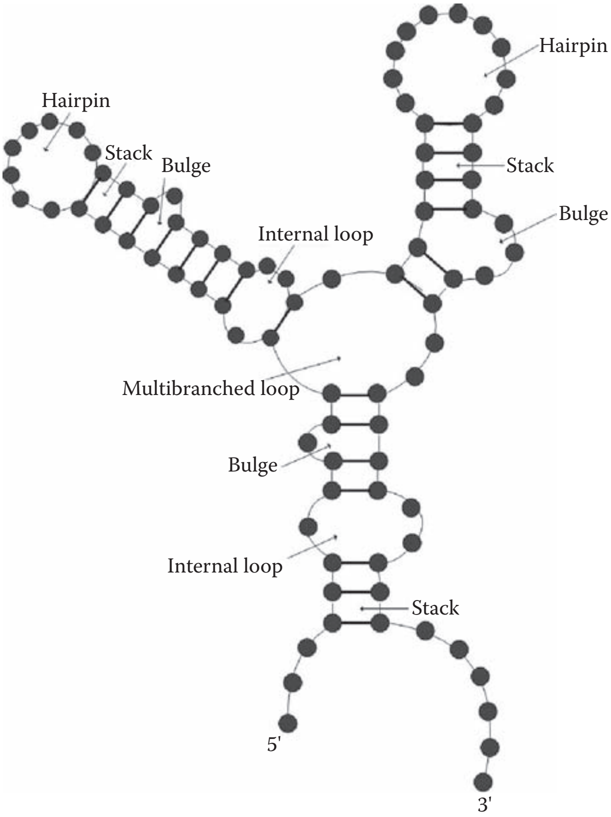


FIGURE 1.1: Illustration of the loops in an RNA secondary structure. Each loop has at least one base pair. A stem consists of two or more consecutive stacks shown in the figure.

the loop components are additive. Our folding algorithm computes the total energy density of R by taking the sum of the energy densities of the loop components in R. Thus, the RNA folding problem can be formalized as follows. Given an RNA sequence S, find the set of base pairs (i, j) and loops with (i, j) as exterior pairs, such that the total energy density of the loops (or equivalently, the exterior pairs) is minimized. The set of base pairs constitutes the optimal secondary structure of S.

When generalizing the folding of a single sequence to the prediction of the consensus structure of a multiple sequence alignment, we introduce the notion of refined alignments. At times, an input alignment may have some columns each of which contains more than 75% gaps. Some tools including RSpredict delete these columns to get a refined alignment [28]; some tools simply use the original input alignment as the refined alignment. Suppose the original input alignment A_o has N sequences and n_o columns, and the refined alignment A has N sequences and n columns, $n \leq n_o$. Formally, the consensus structure of the refined alignment A is a secondary structure R together with its sequence S such that each base pair $(S[i], S[j])$, $1 \leq i < j \leq n$, in R corresponds to the pair of columns i, j in the alignment A, and each base $S[i]$, $1 \leq i \leq n$, is the representative base of the ith column in the alignment A. There are several ways to choose the representative base. For example, $S[i]$ could be the most frequently occurring nucleotide, excluding gaps, in the ith column of the alignment A. Furthermore, there is an energy measure value associated with each base pair $(S[i], S[j])$ or more precisely its corresponding column pair (i, j), such that the total energy measure value of all the base pairs in R is minimized.

The consensus secondary structure of the original input alignment A_o is defined as the structure R_o, obtained from R, as follows: (i) the base (base pair, respectively) for column C_o (column pair (C_o1, C_o2), respectively) in A_o is identical to the base (base pair, respectively) for the corresponding column C (column pair ($C1$, $C2$), respectively) in A if C_o ((C_o1, C_o2), respectively) is not deleted when getting A from A_o; (ii) unpaired gaps are inserted into R, such that each gap corresponds to a column that is deleted when getting A from A_o (see Figure 1.2). In Figure 1.2, the RSpredict algorithm transforms the original input alignment A_o to a refined alignment A by deleting the fourth column (the column in red) of A_o. The algorithm predicts the consensus structure of the refined alignment A. Then the algorithm generates the consensus structure of A_o by inserting an unpaired gap to the fourth position of the consensus structure of A. The numbers inside parentheses in the refined alignment A represent the original column numbers in A_o.

In what follows, we first present an algorithm for folding a single RNA sequence based on the energy density concept described here. We then generalize the algorithm to predict the consensus secondary structure for a set of aligned RNA sequences.

1.2.1.2 Algorithm

The functions and parameters used in our algorithm are defined below where $S[i, j]$ is a subsequence of S and $R[i, j]$ is the optimal secondary structure of $S[i, j]$.

i. NE(i, j) is the total energy density of all loops in $R[i, j]$, where nucleotides at positions i, j may or may not form a base pair.

ii. $\mathrm{NE}_p(i, j)$ is the total energy density of all loops in $R[i, j]$ if nucleotides at positions i, j form a base pair.

iii. $e_H(i, j)(E_H(i, j)$, respectively) is the free energy (energy density, respectively) of the hairpin with exterior pair (i, j).

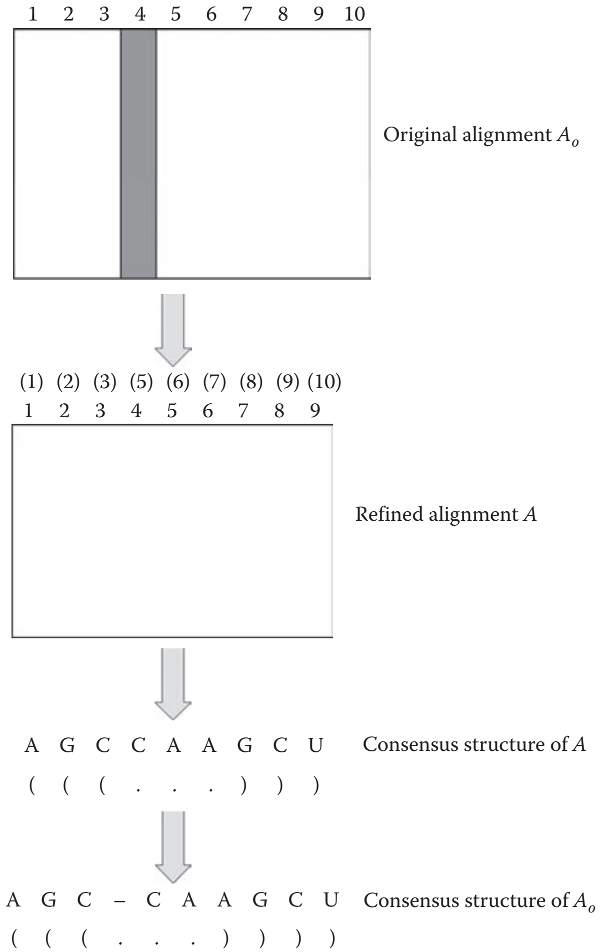


FIGURE 1.2: Illustration of the consensus structure definition used by RSpredict.

iv. $e_S(i,j)(E_S(i,j)$, respectively) is the free energy (energy density, respectively) of the stack with exterior pair (i,j) and interior pair $(i+1,j-1)$.

v. $e_B(i,j,i',\ j')$, $(E_B(i,j,i',\ j')$, respectively) is the free energy (energy density, respectively) of the bulge or internal loop with exterior pair (i,j) and interior pair (i',j').

vi. $e_J\left(i,j,i_1',j_1',i_2',j_2',\ldots,i_k',j_k'\right)$ $E_J\left(i,j,i_1',j_1',i_2',j_2',\ldots,i_k',j_k'\right)$ respectively, is the free energy (energy density, respectively) of the multibranched loop with exterior pair (i,j) and interior pairs $\left(i_1',\ j_1'\right),\ \left(i_2',\ j_2'\right),\ldots,\ \left(i_k',\ j_k'\right)$.

It is clear that

$$E_H\left(i,j\right) = \frac{e_H\left(i,j\right)}{j-i+1} \tag{1.1}$$

$$E_S(i,j) = \frac{e_S(i,j)}{4} \tag{1.2}$$

$$E_B\left(i,j,i',j'\right) = \frac{e_B\left(i,j,i',j'\right)}{i'-i+j-j'+2} \tag{1.3}$$

$$E_J(i,j,i'_1,j'_1,i'_2,j'_2,\ldots,i'_k,j'_k) = \frac{e_J(i,j,i'_1,j'_1,i'_2,j'_2,\ldots,i'_k,j'_k)}{n_1+2\times n_2} \tag{1.4}$$

Here n_1 is the number of unpaired bases and n_2 is the number of base pairs in the multibranched loop in (vi).

Thus, the total energy density of all loops in $R[i,j]$ where (i,j) is a base pair is computed by Equation 1.5:

$$\mathrm{NE}_P\left(i,j\right) = \min \begin{cases} E_H(i,j) \\ E_S\left(i,j\right) + \mathrm{NE}_P\left(i+1,j-1\right) \\ \min\limits_{i<i'<j'<j} \{E_B\left(i,j,i',j'\right) + \mathrm{NE}_P\left(i',j'\right)\} \\ \min\limits_{i<i'_1<j'_1<i'_2<j'_2<\cdots<i'_k<j'_k<j} \{E_J\left(i,j,i'_1,j'_1,i'_2,j'_2,\ldots,i'_k,j'_k\right) \\ \qquad + \sum_{r=1}^{k} \mathrm{NE}_P\left(i'_r,j'_r\right)\} \end{cases} \tag{1.5}$$

That is, the energy density is calculated by taking the minimum of the following four cases:

i. (i,j) is the exterior pair of a hairpin, in which case the energy density $\mathrm{NE}_P(i,\ j)$ equals $E_H(i,j)$, which is the energy density of the hairpin

ii. (i,j) is the exterior pair of a stack, in which case $\mathrm{NE}_P(i,j)$ equals the energy density of the stack, i.e., $E_S(i,j)$, plus $\mathrm{NE}_P(i+1,j-1)$

iii. (i,j) is the exterior pair of a bulge or an internal loop, in which case $\mathrm{NE}_P(i,j)$ equals the minimum of the energy density of the bulge or internal loop $E_B(i,j,i',j')$ plus $\mathrm{NE}_P(i',j')$ for all $i<i'<j'<j$

iv. (i,j) is the exterior pair of a multibranched loop, in which case $\mathrm{NE}_P(i,j)$ equals the minimum of the energy density of the multibranched loop $E_j\left(i,j,i'_1,j'_1,i'_{2,}j'_2,\ldots,i'_k,j'_k\right)$ plus $\sum_{r=1}^{k}\mathrm{NE}_P\left(i'_r,j'_r\right)$, for all $i<i'_1<j'_1<i'_2<j'_2<\cdots<i'_k<j'_k<j$

Equation 1.6 below shows the recurrence formula for calculating $\mathrm{NE}(i,j)$:

$$\mathrm{NE}\left(i,j\right) = \min \begin{cases} \mathrm{NE}\left(i,j-1\right) \\ \mathrm{NE}\left(i+1,j\right) \\ \mathrm{NE}_P\left(i,j\right) \\ \min_{i<h<j}\{\mathrm{NE}\left(i,h-1\right) + \mathrm{NE}\left(h,j\right)\} \end{cases} \tag{1.6}$$

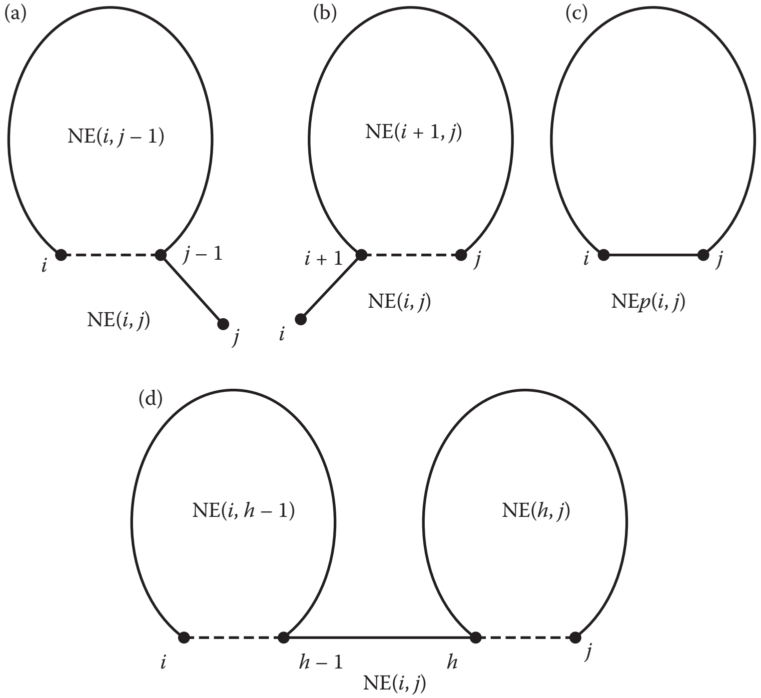


FIGURE 1.3: Illustration of the cases in Equation 1.6. a) the total normalized energy of all loops in the optimal secondary structure $R[i, j-1]$ of subsequence $S[i, j-1]$; b) the total normalized energy of all loops in the optimal secondary structure $R[i+1, j]$ of subsequence $S[i+1, j]$; c) the total normalized energy of all loops in the optimal secondary structure $R[i, j]$ of subsequence $S[i, j]$, where $S[i]$ and $S[j]$ form a base pair; d) the minimum of $\mathrm{NE}(i, k-1)$ plus $\mathrm{NE}(k, j)$ for all $i < k < j$; The dashed line between two nucleotides means that the two nucleotides may or may not form a base pair. The solid line between two nucleotides means that the two nucleotides form a base pair.

That is, the energy density is computed by taking the minimum of the following four cases:

i. The total energy density of all loops in the optimal secondary structure $R\,[i, j-1]$ of subsequence $S\,[i, j-1]$ (Figure 1.3a)

ii. The total energy density of all loops in the optimal secondary structure $R\,[i+1, j]$ of subsequence $S\,[i+1, j]$ (Figure 1.3b)

iii. The total energy density of all loops in the optimal secondary structure $R[i, j]$ of subsequence $S[i, j]$, where $S[i]$ and $S[j]$ form a base pair (Figure 1.3c)

iv. The minimum of NE$(i, h-1)$ plus NE(h, j) for all $i < h < j$ (Figure 1.3d)

Note that case (iii) of Equation 1.6 is not considered when the nucleotides at positions i, j are forbidden to form a base pair, i.e., $(S[i], S[j])$ is a nonstandard base pair. A standard base pair is any of the following: (A,U), (U,A), (G,C), (C,G), (G,U), (U,G); all other base pairs are nonstandard.

In calculating the time complexity of the folding algorithm, there is a need to check for finding the optimal i', j' where $i < i' < j' < j$ in case (iii) (the optimal $i'_1, j'_1, i'_2, j'_2, \ldots, i'_k, j'_k$ where $i < i'_1 < j'_1 < i'_2 < j'_2 < \cdots < i'_k < j'_k < j$ in case (iv), respectively) of Equation 1.5. It can be shown that it takes linear time to compute NE$_P(i, j)$ in Equation 1.5. Hence, the time complexity of the folding algorithm is $O(n^3)$ since we need to calculate NE$_P(i, j)$ for all $1 \le i < j \le n$, where n is the number of nucleotides in the given sequence S. The energy density of the optimal secondary structure R for the sequence S equals NE$(1, n)$.

1.2.2 Calculation of covariance scores

When applying the above folding algorithm to a multiple sequence alignment A_o, we take into consideration the correlation between columns of the alignment. In many cases, the sequences in the alignment may have highly varying lengths. We refine the alignment A_o by deleting columns containing more than 75% gaps to get a refined alignment A [28]. We will use this refined alignment throughout the rest of this subsection.

1.2.2.1 Covariance score

We use the covariance score introduced by RNAalifold [25, 26, 34] to quantify the relationship between two columns in the refined alignment. Let $f_{ij}(XY)$ be the frequency of finding both base X in column i and base Y in column j, where X, Y are in the same row of the refined alignment. We exclude the occurrences of gaps in column i or column j when calculating $f_{ij}(XY)$. The covariation measure for columns i, j, denoted C$_{ij}$, is calculated by Equation 1.7:

$$C_{ij} = \frac{\sum XY, X'Y' f_{ij}(XY)\, D_{ij}(XY, X'Y')\, f_{ij}(X'Y')}{2} \tag{1.7}$$

Here, $D_{ij}(XY, X'Y')$ is the Hamming distance between the two base pairs (X, Y) and (X', Y') if both of the base pairs are standard base pairs, or 0 otherwise. The Hamming distance between (X, Y) and (X', Y') is calculated as follows:

$$D_{ij}(XY, X'Y') = 2 - \delta(X, X') - \delta(Y, Y') \tag{1.8}$$

where

$$\delta(X, X') = \begin{cases} 1 & \text{if } X = X' \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

Observe that the information acquired from the two base pairs (X, Y) and (X', Y') is the same as that from (X', Y') and (X, Y). Thus, we divide the numerator in Equation 1.7 by two so as to obtain the non-redundant information between column i and column j in the refined alignment.

For every pair of columns i, j in the refined alignment, the covariance score of the two columns i and j, denoted Cov_{ij}, is calculated in Equation 1.10:

$$\mathrm{Cov}_{ij} = C_{ij} + c_1 \times \mathrm{NF}_{ij} \tag{1.10}$$

Here, C_{ij} is as defined in Equation 1.7, c_1 is a user-defined coefficient (in the study presented here, c_1 has a value of -1), and

$$\mathrm{NF}_{ij} = \frac{\mathrm{NC}_{ij}}{N} \tag{1.11}$$

where N is the total number of sequences and NC_{ij} is the total number of conflicting sequences in the refined alignment. A conflicting sequence is one that has a gap in column i or column j, or has a nonstandard base pair in the columns i, j of the refined alignment. A sequence with gaps in both columns i, j is not conflicting.

1.2.2.2 Pairing threshold

We say that column i and column j in the refined alignment can possibly form a base pair if their covariance score is greater than or equal to a pairing threshold; otherwise, column i and column j are forbidden to form a base pair. The pairing threshold, η, used in RSpredict is calculated as follows.

It is known that, on average, 54% of the nucleotides in an RNA sequence S are involved in the base pairs of its secondary structure [35]. We use this information to calculate an alignment-dependent pairing threshold, observing that the base pairs in the consensus secondary structure of a sequence alignment represent the column pairs with the highest covariance scores. Given that different structures contain different numbers of base pairs, we consider two different percentages of columns, namely, 30% and 65%, in the sequence alignment. For each percentage p, there are at most T_p possible base pairs, where

$$T_p = \frac{(p \times n) \times (p \times n - 1)}{2} \tag{1.12}$$

and n is the number of columns in the sequence alignment.

Now, we calculate the covariance scores of all pairs of columns in the given refined alignment, and sort the covariance scores in descending order. We then select the top T_p largest covariance scores and store the covariance scores in the set ST_p. Thus, the set $\mathrm{ST}_{0.65}$ contains the top largest covariance scores that involve 65% of the columns in the refined alignment; the set $\mathrm{ST}_{0.30}$ contains the top largest covariance scores that involve 30% of the columns in the refined alignment; and $\mathrm{ST}_{0.65} \backslash \mathrm{ST}_{0.30}$ is the set difference that contains covariance scores in $\mathrm{ST}_{0.65}$ but not in $\mathrm{ST}_{0.30}$ (see Figure 1.4). The pairing

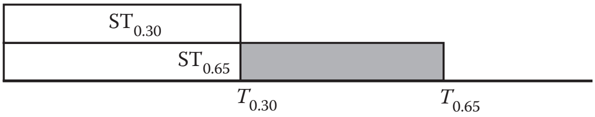


FIGURE 1.4: Illustration of the pairing threshold computation. The pairing threshold used in RSpredict is computed as the average of the covariance scores inside the shaded area.

threshold η used in RSpredict is calculated as the average of the covariance scores in $\mathrm{ST}_{0.65} \backslash \mathrm{ST}_{0.30}$, as shown in Equation 1.13:

$$\eta = \frac{\sum \mathrm{Cov}_{ij} \in \mathrm{ST}_{0.65} \backslash \mathrm{ST}_{0.30} \mathrm{Cov}_{ij}}{|\mathrm{ST}_{0.65} \backslash \mathrm{ST}_{0.30}|} \tag{1.13}$$

where the denominator is the cardinality of the set difference $\mathrm{ST}_{0.65} \backslash \mathrm{ST}_{0.30}$.

If the covariance score of columns i and j is greater than or equal to η, then column i and column j can possibly form a base pair, and we refer to (i, j) as a pairing column. If the covariance score of the columns i and j is less than η, we will check the covariance scores of the immediate neighboring column pairs of i, j to see if they are above a user-defined threshold [31] (in the study presented here, this threshold is set to 0). The immediate neighboring column pairs of i, j are $i + 1$, $j - 1$ and $i - 1$, $j + 1$. If the covariance scores of both of the immediate neighboring column pairs of i, j are greater than or equal to $\max\{\eta, 0\}$, then (i, j) is still considered as a paring column.

1.2.3 Algorithms for RSpredict

Given a refined multiple sequence alignment A with N sequences, let (i, j) be a pairing column in A. Let X_i^S (Y_j^S, respectively) be the nucleotide at position i (j, respectively) of the sequence S in the alignment A. $\left(X_i^S, Y_j^S\right)$ must be the exterior pair of some loop L in S. We use $e\left(X_i^S, Y_j^S\right)$ to represent the free energy of that loop L. If $\left(X_i^S, Y_j^S\right)$ is a nonstandard base pair, $e\left(X_i^S, Y_j^S\right) = 0$. We assign the pairing column (i, j) a pseudo-energy e_{ij} where

$$e_{ij} = \frac{1}{N} \sum_{S \in A} e\left(X_i^S, Y_j^S\right) + c_2 \times \mathrm{Cov}_{ij} \tag{1.14}$$

Here, c_2 is a user-defined coefficient (in the study presented here, $c_2 = -1$). Thus, every pairing column in the refined alignment A has a pseudo-energy. We then apply the minimum energy density folding algorithm described in the beginning of this section to the refined alignment A, treating each pairing column in A as a possible base pair considered in the folding algorithm.

Notice that when calculating the energy density for the loop L, the sequence S is in the refined alignment A, which may have fewer columns than the original input alignment A_o (cf. Figure 1.2). RSpredict computes all energy densities based on the refined alignment, and the program uses loop lengths from the refined alignment A rather than the original input alignment A_o. Let R be the consensus secondary structure, computed by RSpredict, for the refined alignment A. We obtain the consensus structure R_o of the original input alignment A_o by inserting unpaired gaps to the positions in R whose corresponding columns are deleted when getting A from A_o (cf. Figure 1.2). The following summarizes the algorithms for RSpredict:

1. Input an alignment A_o in the FASTA or ClustalW format.
2. Delete the columns with more than 75% gaps from A_o to obtain a refined alignment A.
3. Compute the pseudo-energy e_{ij} for every pairing column (i, j) in A as in Equation 1.14.
4. Run the minimum energy density folding algorithm on A, using the pseudo-energy values obtained from step (3) to produce the consensus secondary structure R of the refined alignment A. The base at position i of the consensus secondary structure R is the most frequently occurring nucleotide, excluding gaps, in the ith column of the refined alignment A.
5. Map the consensus structure R back to the original alignment A_o by inserting unpaired gaps to the positions of R whose corresponding columns are deleted in Step (2).

Notice that Equation 1.6 is used to compute the NE values only. To generate the optimal structure R in Step (4), we maintain a stack of pointers that point to the substructures of loops with minimum energy density as we compute the NE values. Once all the NE values are calculated and the energy density of the optimal secondary structure R is obtained, we pop up the pointers from the stack to extract the optimal predicted structure. In step (5), we map the bases (base pairs, respectively) for the columns (column pairs, respectively) in A to their corresponding columns (column pairs, respectively) in A_o. For example, consider Figure 1.2 again. In the figure, the refined alignment A is obtained by deleting column 4 from the original input alignment A_o. The bases for columns 1, 2, 3, 4 in A are mapped to columns 1, 2, 3, 5 in A_o. The base pair between column 1 and column 9 in A becomes the base pair between column 1 and column 10 in A_o; the base pair between column 2 and column 8 in A becomes the base pair between column 2 and column 9 in A_o. An unpaired gap is inserted to the position corresponding to the deleted column 4 in A_o.

Let N be the number of sequences and n_o be the number of columns in the input alignment A_o. Step (2) takes $O(Nn_o)$ time. Step (3) takes $O\left(n_o^2\right)$ time. Step (4) takes $O\left(n_o^3\right)$ time. Step (5) takes $O(n_o)$ time. Therefore, the time complexity of RSpredict is $O\left(Nn_o + n_o^3\right)$, which is approximately $O\left(n_o^3\right)$ as $N$is usually much smaller than n_o.

1.3 Results

We conducted a series of experiments to evaluate the performance of RSpredict and compared it with five related tools including KNetFold, Pfold, RNAalifold, RSefold, and RSdfold. We tested these tools on Rfam [36] sequence alignments with different similarities. The Rfam sequence alignments come with consensus structures. For evaluation purposes, we used the Rfam consensus structures as reference structures and compared them against the consensus structures predicted by the six tools. The similarity of a sequence alignment is determined by the average pairwise sequence identity (APSI) of that alignment [6]. In the study presented here, a sequence alignment is of high similarity if its APSI value is greater than 75%, is of medium similarity if its APSI value is between 55% and 75%, or is of low similarity if its APSI value is less than 55%. The data sets used in testing included 20 Rfam sequence alignments of high similarity and 36 Rfam sequence alignments of low and medium similarity. These data sets were chosen to form a collection of sequence alignments with different (low, medium and high) APSI values, different numbers of sequences, as well as different sequence alignment lengths. More specifically, the data sets contained sequence alignments that ranged in size from 2 to 160 sequences, in length from 33 to 262 nucleotides and had APSI values ranging from 42% to 99%.

The performance measures used in our study include sensitivity (SN) and selectivity (SL) [6], where

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1.15}$$

$$\mathrm{SL} = \frac{\mathrm{TP}}{\mathrm{TP} + (\mathrm{FP} - \xi)}. \tag{1.16}$$

Here, TP is the number of correctly predicted base pairs ("true positives"), FN is the number of base pairs in a reference structure that were not predicted ("false negatives") and FP is the number of incorrectly predicted base pairs ("false positives"). False positives are classified as inconsistent, contradicting or compatible [6]. When predicting the consensus secondary structure for a multiple sequence alignment, a predicted base pair (i, j) is inconsistent if column i in the alignment is paired with column $q, q \neq j$, or column j is paired with column $p, p \neq i$, and p, q form a base pair in the reference structure of the alignment. A base pair (i, j) is contradicting if there exists a base pair (p, q) in the reference structure of the alignment, such that $i < p < j < q$. A base pair (i, j) is compatible if it is a false positive but is neither inconsistent nor contradicting. The ξ in SL represents the number of compatible base pairs, which are considered neutral with respect to algorithmic accuracy. Therefore ξ is subtracted from FP. Finally, we used the Matthews correlation coefficient (MCC) to combine the sensitivity and selectivity, where MCC is approximated to the geometric mean of the two measures, i.e., MCC $\approx \sqrt{\mathrm{SN} \times \mathrm{SL}}$ [18]. The larger MCC, SN, SL values a tool has, the better performance that tool achieves and the more accurate that tool is.

1.3.1 Performance evaluation on Rfam alignments of high similarity

The first data set consisted of seed alignments of high similarity taken from 20 families in Rfam. The APSI values of these seed alignments ranged from 77% to 99%. The alignments ranged in size from 2 to 160 sequences and in length from 33 to 159 nucleotides. Table 1.1 presents the accession number, description, number of sequences, and length of the seed alignment of each of the 20 Rfam families used in the experiment. The seed alignments of the 20 families are of high similarity; their APSI values are shown in the last column of the table. The families are sorted, from top to bottom, in ascending order on the APSI values. All six tools including RSpredict, KNetFold, RNAalifold, Pfold, RSefold and RSdfold were tested on this data set.

The graphs in Figure 1.5 show the trend of the MCC, SN, and SL, which are sorted in descending order for each tool under analysis. The X-axis shows, therefore, the rank of the MCC (SN and SL, respectively) from highest to lowest. For example, number 1 in the X-axis corresponds to the highest score achieved by each tool. The Y-axis represents the MCC, SN, and SL, respectively.

It can be seen from Figure 1.5 that RSpredict performed the best while RSdfold performed the worst among the six tools. The Pfold tool had good performance in selectivity but did not perform well in sensitivity and as a result in MCC. It also suffered from a size limitation (the Pfold web server can accept a multiple alignment of up to 40 sequences). Only 17 out of the 20 sequence alignments used in the experiment were accepted by the Pfold server; the other three alignments (RF00386, RF00041, and RF00389) had more than 40 sequences and therefore could not be run on the Pfold server. RSpredict had stable performance with the best mean 0.85 (standard deviation 0.16, respectively) in MCC, while the other methods' MCC values varied a lot and had means (standard deviations, respectively) ranging from 0.37 to 0.82 (0.24 to 0.34, respectively).

1.3.2 Performance evaluation on Rfam alignments of medium and low similarity

In the second experiment, we compared RSpredict with the other five methods on multiple sequence alignments of low and medium similarity. The test dataset included seed alignments of 36 families taken from Rfam [36]. The APSI values of the seed alignments ranged from 42 to 75%, the number of sequences in the alignments ranged from 3 to 114, and the alignment lengths ranged from 43 to 262 nucleotides. Table 1.2 presents the accession number,

TABLE 1.1: Rfam alignments of high similarity.

Accession	Description	Number of sequences	Length	APSI
RF00460	U1A polyadenylation inhibition element (PIE)	8	75	77%
RF00326	Small nucleolar RNA Z155	8	81	79%
RF00560	Small nucleolar RNA SNORA17	38	132	82%
RF00453	Cardiovirus cis-acting replication element (CRE)	12	33	82%
RF00386	Enterovirus 5′ cloverleaf cis-acting replication element	160	91	83%
RF00421	Small nucleolar RNA SNORA32	9	122	84%
RF00302	Small nucleolar RNA SNORA65	8	130	84%
RF00465	Japanese encephalitis virus (JEV) hairpin structure	20	60	86%
RF00501	Rotavirus cis-acting replication element (CRE)	14	68	87%
RF00041	Enteroviral 3′ UTR element	60	123	87%
RF00575	Small nucleolar RNA SNORD70	4	88	89%
RF00362	Pospiviroid RY motif stem loop	16	79	92%
RF00105	Small nucleolar RNA SNORD115	23	82	92%
RF00467	Rous sarcoma virus (RSV) primer binding site (PBS)	23	75	93%
RF00389	Bamboo mosaic virus satellite RNA cis-regulatory element	42	159	93%
RF00384	Poxvirus AX element late mRNA cis-regulatory element	7	62	93%
RF00098	Snake H/ACA box small nucleolar RNA	22	150	93%
RF00607	Small nucleolar RNA SNORD98	2	67	98%
RF00320	Small nucleolar RNA Z185	2	86	98%
RF00318	Small nucleolar RNA Z175	3	81	99%

description, number of sequences, and length of the seed alignment of each of the 36 Rfam families used in the experiment. The seed alignments of the 36 families are of low and medium similarity; their APSI values are shown in the last column of the table. The families are sorted, from top to bottom, in ascending order on the APSI values.

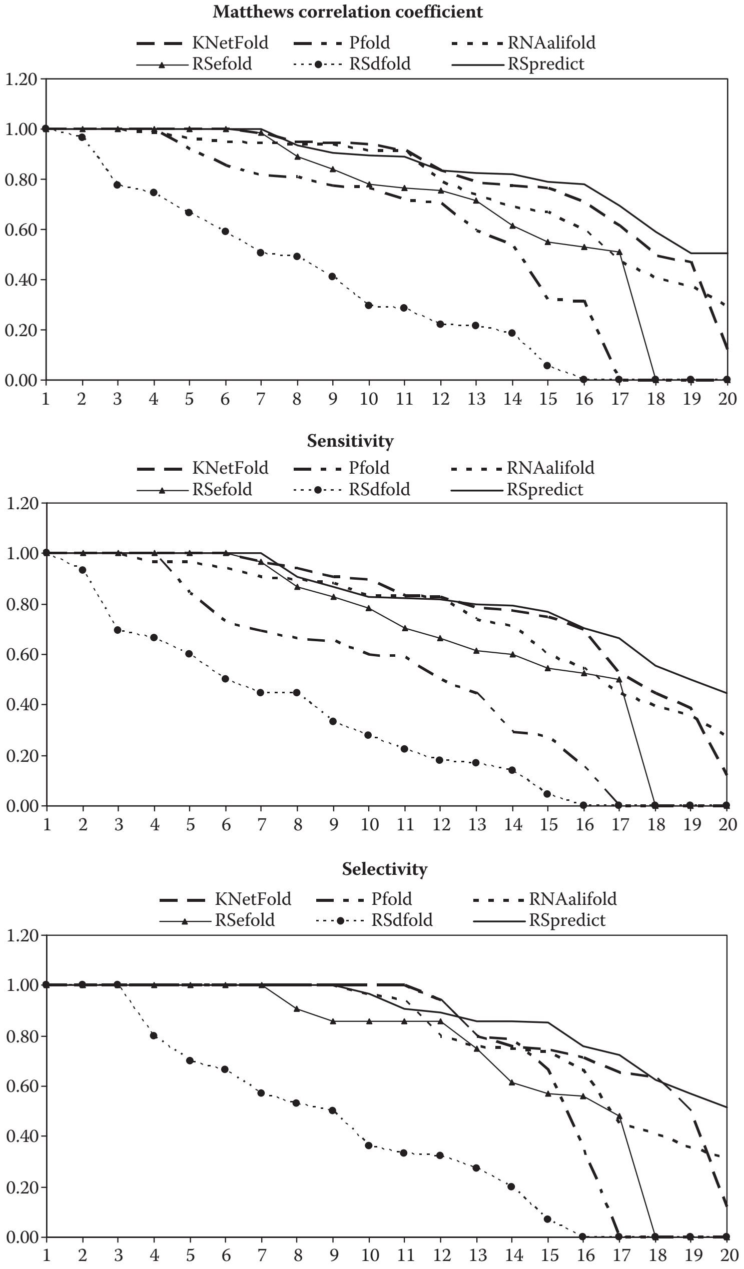
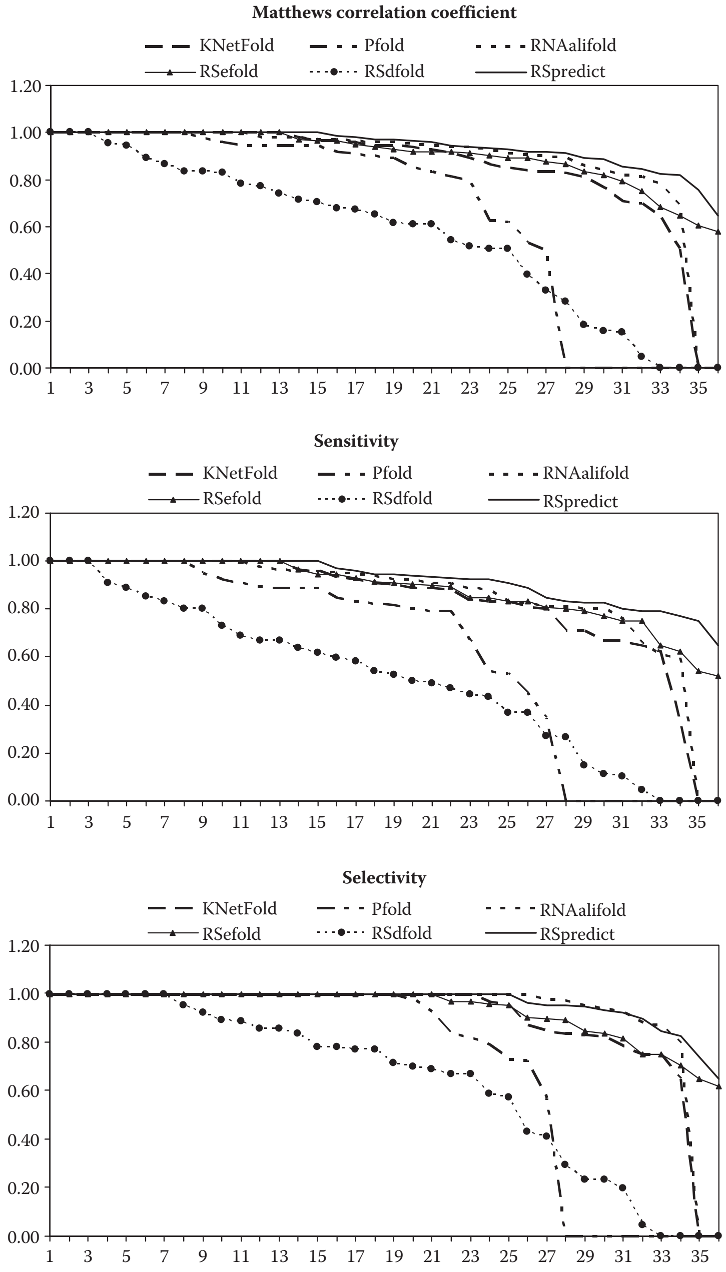


FIGURE 1.5: Comparison of the MCC, SN, and SL values of the six tools under analysis on the seed alignments of high similarity taken from the 20 families listed in Table 1.1.

TABLE 1.2: Rfam alignments of low and medium similarity.

Accession	Description	Number of sequences	Length	APSI
RF00230	T-box leader	103	262	42%
RF00080	yybP-ykoY leader	50	131	44%
RF00515	PyrR binding site	72	125	47%
RF00557	Ribosomal protein L10 leader	66	149	48%
RF00504	Glycine riboswitch	93	111	50%
RF00029	Group II catalytic intron	114	94	52%
RF00458	Cripavirus internal ribosome entry site (IRES)	7	203	54%
RF00559	Ribosomal protein L21 leader	33	81	54%
RF00234	glmS glucosamine-6-phosphate activated ribozyme	11	218	55%
RF00556	Ribosomal protein L19 leader	24	43	55%
RF00519	suhB	13	80	56%
RF00379	ydaO/yuaA leader	25	150	58%
RF00380	ykoK leader	36	172	59%
RF00445	mir-399 microRNA precursor family	13	119	59%
RF00522	PreQ1 riboswitch	22	47	59%
RF00095	Pyrococcus C/D box small nucleolar RNA	25	59	60%
RF00442	ykkC-yxkD leader	11	111	60%
RF00430	Small nucleolar RNA SNORA54	5	134	60%
RF00521	SAM riboswitch (alpha-proteobacteria)	12	79	61%
RF00049	Small nucleolar RNA SNORD36	20	82	63%
RF00513	Tryptophan operon leader	11	100	63%
RF00309	Small nucleolar RNA snR60/ Z15/Z230/Z193/J17	23	106	63%
RF00451	mir-395 microRNA precursor family	21	112	64%
RF00464	mir-92 microRNA precursor family	33	80	64%
RF00507	Coronavirus frameshifting stimulation element	23	85	66%
RF00388	Qa RNA	5	103	70%
RF00357	Small nucleolar RNA R44/ J54/Z268 family	19	105	70%
RF00434	Luteovirus cap-independent translation element (BTE)	17	108	71%
RF00525	Flavivirus DB element	111	76	71%
RF00581	Small nucleolar SNORD12/ SNORD106	8	91	71%
RF00238	ctRNA	48	88	72%
RF00477	Small nucleolar RNA snR66	5	105	72%
RF00608	Small nucleolar RNA SNORD99	3	80	72%
RF00468	Heaptitis C virus stem-loop VII	110	66	74%
RF00489	ctRNA	14	80	74%
RF00113	QUAD RNA	14	150	75%

The MCC, SN, and SL values are sorted in descending order for each tool under analysis and placed in the graphs in Figure 1.6. The X-axis shows, therefore, the rank of the MCC (SN and SL, respectively) from highest to lowest. For example, number 1 in the X-axis corresponds to the highest score achieved by each tool. The Y-axis represents the MCC, SN, and SL, respectively.


Matthews correlation coefficient

Sensitivity

Selectivity

FIGURE 1.6: Comparison of the MCC, SN, and SL values of the six tools under analysis on the seed alignments of low and medium similarity taken from the 36 families listed in Table 1.2.

Comparing Figures 1.5 and 1.6, we see that the methods under analysis generally performed better on sequence alignments of medium and low similarity than on sequence alignments of high similarity. Like what was observed in the previous experiment, RSdfold performed the worst (cf. Figure 1.5). The structures predicted by RSdfold tend to be stem-like structures; therefore, many structures, particularly those containing multibranched loops, were mispredicted. For this reason, RSdfold yielded very low MCC, SN and SL values.

RSpredict outperformed the other five methods based on the three performance measures used in the experiment. The tool achieved a high mean value of 0.94 in MCC, better than those of KNetFold (0.86), Pfold (0.88) and RNAalifold (0.89). Similar results were observed for sensitivity and selectivity values. Furthermore, RSpredict exhibited stable performance across all the families tested in the experiment. The tool had an MCC, SN and SL standard deviation of 0.08, 0.09 and 0.08, respectively. These numbers were better than the standard deviation values obtained from the other five methods, which ranged from 0.11 to 0.34. Pfold suffered from a size limitation; it could not generate a structure for the large seed alignments with more than 40 sequences in 9 families, including RF00230, RF00080, RF00515, RF00557, RF00504, RF00029, RF00525, RF00238 and RF00468.

1.4 Conclusions

In this chapter we presented a software tool, called RSpredict, capable of predicting the consensus secondary structure for a set of aligned RNA sequences via energy density minimization and covariance score calculation. Our experimental results showed that RSpredict is competitive with some widely used tools including RNAalifold and Pfold on tested datasets, suggesting that RSpredict can be a choice when biologists need to predict RNA secondary structures of multiple sequence alignments, especially those with low and medium similarity. Notice that RSpredict differs from KNetFold [31] in that KNetFold is a machine learning method that relies on precompiled training data derived from existing RNA secondary structures. RSpredict, on the other hand, is based on a dynamic programming algorithm for folding sequences and does not utilize training data.

Given a multiple sequence alignment A_o, our work is focused on predicting the consensus structure of the aligned sequences in A_o, rather than folding each individual sequence in A_o. Our approach is to first transform A_o to a refined alignment A by deleting columns with more than 75% gaps from A_o, then predict the consensus structure for A, and finally extend the consensus structure by inserting gaps to the positions corresponding to the deleted columns in A_o (cf. Figure 1.2). The predicted structure may not correspond exactly to any individual sequence in the original alignment A_o. As an example, assume for simplicity that A_o is the same as A, i.e., no columns are deleted when getting A from A_o. Consider a particular sequence S in A_o. Assume that the position (column) i of S has a gap due to the alignment with the other sequences in A_o. On the other hand, the position i in the consensus structure of A_o has the most frequently occurring nucleotide in column i of A_o, which cannot be a gap. As a result, the consensus structure of A_o, which is at least one nucleotide longer than S, cannot be mapped exactly back onto S. In future work we plan to look into ways for improving on consensus structure prediction. Possible ways include the utilization of evolutionary information [37], more sophisticated models of covariance scoring, and training data for more accurate pairing thresholds.

References

[1] Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.

[2] Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* 31:3429–3431.

[3] Shapiro, B.A., Kasprzak, W., Grunewald, C., Aman, J. 2006. Graphical exploratory data analysis of RNA secondary structure dynamics predicted by the massively parallel genetic algorithm. *J. Mol. Graph. Model.* 25:514–531.

[4] Bellamy-Royds, A.B., Turcotte, M. 2007. Can Clustal-style progressive pairwise alignment of multiple sequences be used in RNA secondary structure prediction? *BMC Bioinformatics* 8:190.

[5] Horesh, Y., Doniger, T., Michaeli, S., Unger, R. RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. *BMC Bioinformatics* 8:366.

[6] Gardner, P.P., Giegerich, R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 5:140.

[7] Alkan, C., Karakoc, E., Sahinalp, S.C., Unrau, P., Alexander, E., Zhang, K., Buhler, J. 2006. RNA secondary structure prediction via energy density minimization. In *Proceedings of the Research in Computational Molecular Biology* (*RECOMB*), Springer Berlin/Heidelberg, Venice, Italy, 130–142.

[8] Xia, T., SantaLucia, J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., Turner, D.H. 1998. Thermodynamic parameters for an

expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* 37:14719–14735.

[9] Xu, X., Yongmei, J., Stormo, G.D. 2007. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics* 23:1883–1891.

[10] Giegerich, R., Voss, B., Rehmsmeier, M. 2007. Abstract shapes of RNA. *Nucleic Acids Res.* 32:4843–4851.

[11] Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., Giegerich, R. 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 22:500–503.

[12] Siebert, S., Backofen, R. 2005. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 21:3352–3359.

[13] Shapiro, B.A., Bengali, D., Kasprzak, W., Wu, J.C. 2001. RNA folding pathway functional intermediates: their prediction and analysis. *J. Mol. Biol.* 312:27–44.

[14] Khaladkar, M., Bellofatto, V., Wang, J.T.L., Tian, B., Shapiro, B.A. 2007. RADAR: a web server for RNA data analysis and research. *Nucleic Acids Res.* 35:W300–W304.

[15] Liu, J., Wang, J.T.L., Hu, J., Tian, B. 2005. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics* 6:89.

[16] Ji, Y., Xu, X., Stormo, G.D. 2004. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics* 20:1591–1602.

[17] Bafna, V., Tang, H., Zhang, S. 2006. Consensus folding of unaligned RNA sequences revisited. *J. Comput. Biol.* 13:283–295.

[18] Gorodkin, J., Stricklin, S.L., Stormo, G.D. 2001. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.* 29:2135–2144.

[19] Mathews, D.H., Turner, D.H. 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* 317:191–203.

[20] Holmes, I., Rubin, G.M. 2002. Pairwise RNA structure comparison with stochastic context-free grammars. In *Proceedings of the Pacific Symposium Biocomputing*, Lihue, Hawaii, 163–174.

[21] Hofacker, I.L., Bernhart, S.H.F., Stadler, P.F. 2004. Alignment of RNA base pairing probability matrices. *Bioinformatics* 20:2222–2227.

[22] Lindgreen, S., Gardner, P.P., Krogh, A. 2007. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* 23:3304–3311.

[23] Touzet, H., Perriquet, O. 2004. CARNAC: folding families of related RNAs. *Nucleic Acids Res.* 32:W142–W145.

[24] Sankoff, D. 1985. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* 45:810–825.

[25] Hofacker, I.L., Fekete, M., Stadler, P.F. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* 319:1059–1066.

[26] Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., Stadler, P.F. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9:474.

[27] Klein, R.J., Eddy, S.R. 2003. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics* 4:44.

[28] Knudsen, B., Hein, J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* 31:3423–3428.

[29] Cary, R.B., Stormo, G.D. 1995. Graph-theoretic approach to RNA modeling using comparative data. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, 75–80.

[30] Tabaska, J.E., Cary, R.B., Gabow, H.N., Stormo, G.D. 1998. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics* 14:691–699.

[31] Bindewald, E., Shapiro, B.A. 2006. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* 12:342–352.

[32] Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., Turner, D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA.* 101:7287–7292.

[33] Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.

[34] Lindgreen, S., Gardner, P.P., Krogh, A. 2006. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics* 22:2988–2995.

[35] Mathews, D.H., Banerjee, A.R., Luan, D.D., Eickbush, T.H., Turner, D.H. 1997. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* 3:1–16.

[36] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R. 2003. Rfam: an RNA family database. *Nucleic Acids Res.* 31:439–441.

[37] Seemann, S.E., Gorodkin, J., Backofen, R. 2008. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.* 36:6355–6362.

References

1 Chapter 1. Consensus Structure Prediction for RNA Alignments

[1] Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406–3415.

[2] Hofacker, I.L. 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31:3429–3431.

[3] Shapiro, B.A., Kasprzak, W., Grunewald, C., Aman, J. 2006. Graphical exploratory data analysis of RNA secondary structure dynamics predicted by the massively parallel genetic algorithm. J. Mol. Graph. Model. 25:514–531.

[4] Bellamy-Royds, A.B., Turcotte, M. 2007. Can Clustal-style progressive pairwise alignment of multiple sequences be used in RNA secondary structure prediction? BMC Bioinformatics 8:190.

[5] Horesh, Y., Doniger, T., Michaeli, S., Unger, R. RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. BMC Bioinformatics 8:366.

[6] Gardner, P.P., Giegerich, R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5:140.

[7] Alkan, C., Karakoc, E., Sahinalp, S.C., Unrau, P., Alexander, E., Zhang, K., Buhler, J. 2006. RNA secondary structure prediction via energy density minimization. In Proceedings of the Research in Computational Molecular Biology (RECOMB), Springer Berlin/Heidelberg, Venice, Italy, 130–142.

[8] Xia, T., SantaLucia, J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., Turner, D.H. 1998. Thermodynamic parameters for an T&F Cat # C6847 Chapter: 1 page: 24 date: August 5, 2009 expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37:14719–14735.

[9] Xu, X., Yongmei, J., Stormo, G.D. 2007. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. Bioinformatics 23:1883–1891.

[10] Giegerich, R., Voss, B., Rehmsmeier, M. 2007. Abstract shapes of RNA. Nucleic Acids Res. 32:4843–4851.

[11] Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., Giegerich, R. 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics 22:500–503.

[12] Siebert, S., Backofen, R. 2005. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics 21:3352–3359.

[13] Shapiro, B.A., Bengali, D., Kasprzak, W., Wu, J.C. 2001. RNA folding pathway functional intermediates: their prediction and analysis. J. Mol. Biol. 312:27–44.

[14] Khaladkar, M., Bellofatto, V., Wang, J.T.L., Tian, B., Shapiro, B.A. 2007. RADAR: a web server for RNA data analysis and research. Nucleic Acids Res. 35:W300–W304.

[15] Liu, J., Wang, J.T.L., Hu, J., Tian, B. 2005. A method for aligning RNA secondary structures and its application to RNA motif detection. BMC Bioinformatics 6:89.

[16] Ji, Y., Xu, X., Stormo, G.D. 2004. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. Bioinformatics 20:1591–1602.

[17] Bafna, V., Tang, H., Zhang, S. 2006. Consensus folding of unaligned RNA sequences revisited. J. Comput. Biol. 13:283–295.

[18] Gorodkin, J., Stricklin, S.L., Stormo, G.D. 2001. Discovering common stem-loop motifs in unaligned RNA sequences. Nucleic Acids Res. 29:2135–2144.

[19] Mathews, D.H., Turner, D.H. 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J. Mol. Biol. 317:191–203.

[20] Holmes, I., Rubin, G.M. 2002. Pairwise RNA structure comparison with stochastic context-free grammars. In Proceedings of the Pacific Symposium Biocomputing, Lihue, Hawaii, 163–174. T&F Cat # C6847 Chapter: 1 page: 25 date: August 5, 2009

[21] Hofacker, I.L., Bernhart, S.H.F., Stadler, P.F. 2004. Alignment of RNA base pairing probability matrices.

Bioinformatics 20:2222-2227.

[22] Lindgreen, S., Gardner, P.P., Krogh, A. 2007. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. Bioinformatics 23:3304-3311.

[23] Touzet, H., Perriquet, O. 2004. CARNAC: folding families of related RNAs. Nucleic Acids Res. 32:W142-W145.

[24] Sankoff, D. 1985. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. 45:810-825.

[25] Hofacker, I.L., Fekete, M., Stadler, P.F. 2002. Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. 319:1059-1066.

[26] Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., Stadler, P.F. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474.

[27] Klein, R.J., Eddy, S.R. 2003. RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 4:44.

[28] Knudsen, B., Hein, J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 31:3423-3428.

[29] Cary, R.B., Stormo, G.D. 1995. Graph-theoretic approach to RNA modeling using comparative data. In Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, 75-80.

[30] Tabaska, J.E., Cary, R.B., Gabow, H.N., Stormo, G.D. 1998. An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics 14:691-699.

[31] Bindewald, E., Shapiro, B.A. 2006. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA 12:342-352.

[32] Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., Turner, D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. USA. 101:7287-7292.

[33] Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J. Mol. Biol. 288:911–940.

[34] Lindgreen, S., Gardner, P.P., Krogh, A. 2006. Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformatics 22:2988–2995.
T&F Cat # C6847 Chapter: 1 page: 26 date: August 5, 2009

[35] Mathews, D.H., Banerjee, A.R., Luan, D.D., Eickbush, T.H., Turner, D.H. 1997. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. RNA 3:1–16.

[36] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R. 2003. Rfam: an RNA family database. Nucleic Acids Res. 31:439–441.

[37] Seemann, S.E., Gorodkin, J., Backofen, R. 2008. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Res. 36:6355–6362.

2 Chapter 2. Invariant Geometric Properties of Secondary Structure Elements in Proteins

Agrawal, R., Imielinski, T., and Swami, A.N., 1993. Mining association rules between sets of items in large databases. Proc. of the ACM SIGMOD Intl. Conference on Management of Data, 207–216.

Comin, M., Guerra, C., and Zanotti, G. 2008. Mining over-represented 3D patterns of secondary structures in proteins. Journal of Bioinformatics and Computational Biology, 6(6):1067–1087.

Comin, M., Guerra, C., and Zanotti, G. 2004. PROUST: A comparison method of three-dimensional structures of proteins using indexing techniques. Journal of Computational Biology, 11(6):1061–1072.

Dror, O., Benyamini, H., Nussinov, R., and Wolfson, H. 2003. Multiple structural alignment by secondary structures: algorithm and applications. Protein Science, 12:2492–2507.

Gerstein, M., and Hegyi, H. 1998. Comparing genomes in terms of protein structures: surveys of a finite parts list. FEMS Microbiology, 22:277–304.

Gibrat, J.-F., Madej, T., and Bryant, S.H. 1996. Surprising similarities in structure comparison. Current Opinion in Structural Biology, 6:377–385.

Guerra, C., Lonardi, S., and Zanotti, G. 2002. Analysis of secondary structures using indexing techniques. IEEE Proc. First Int. Symposium on 3D Data Processing Visualization and Transmission, 812–821.

Holm, L., and Sander, C. 1996. Mapping the protein universe. Science, 273:595–602.

Horn, B.K.P. 1987. Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America, 4(4):629–642. T&F Cat # C6847 Chapter: 2 page: 48 date: August 5, 2009

Kolodny, R., Koehl, P., and Levitt, M. 2005. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. Journal of Molecular Biology, 346:1173–1188.

Murzin, A., Brenner, S.E., Hubbard, T., and Chotia, C. 1995. SCOP: a structural classification of proteins for the investigation of sequences and structures. Journal of Molecular Biology, 247:536-540.

Orengo, C.A., and Thornton, J.M. 2005. Protein families and their evolution— A structural perspective. Annual Review of Biochemistry, 74:867-900.

Platt, D.E., Guerra, C., Zanotti, G., and Rigotsous, I. 2003. Global secondary structure packing angle bias in proteins. Proteins: Structure, Functions, and Genetics, 53:252-261.

Shatsky, M., Nussinov, R., and Wolfson, H.J. 2004. A method for simultaneous alignment of multiple protein structures. Proteins, 156(1):143-156.

Shatsky, M., Nussinov, R., and Wolfson, H.J. 2006. Optimization of multiplesequence alignment based on multiple-structure alignment. Proteins: Structure, Functions Bioinformatics, 62:209-217.

Shindyalov, I.N., and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering, 11(9):739-747.

Singh, A.P., and Brutlag, D.L. 1997. Hierarchical protein structure superposition using both secondary structures and atomic representations. Proc. 5th Int. Conf. Intell. Sys. Mol. Biology, 284-293.

Wang, X., Jason, T.L., Shasha D., Shapiro, B.A., Rigoutsos, I., and Zhang, K. 2002. Finding patterns in three-dimensional graphs: algorithms and applications to scientific data mining. IEEE Transactions on Knowledge and Data Mining, 14(4):731-749.

3 Chapter 3. Discovering 3D Motifs in RNA

Apostolico, A., Ciriello, G., Guerra, C., Heitsch, C.E., Hsiao, C., and Williams, L.D. 2009. Finding 3D motifs in ribosomal RNA structures. Nucleic Acids Research, doi: 10.1093/nar/gkn1044.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science, 289, 905–920.

Bindewald, E., Hayes, R., Yingling, Y.G., Kasprzak, W., and Shapiro, B.A. 2008. RNA Junction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. Nucleic Acids Research, 36, D392-D397, doi: 10.1093/nar/gkm842.

Burks, J., Zwieb, C., Muller, F., Wower, I., and Wower, J. 2005. Comparative 3-D modeling of TmRNA. BMC Molecular Biology, 6, 14, doi: 10.118611471-2199-6-14.

Dror, O., Nussinov, R. and Wolfson, H. 2005. ARTS: alignment of RNA tertiary structures. Bioinformatics, 21(37), 47–53.

Duarte, C.M., Wadley, L.M., and Pyle, A.M. 2003. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. Nucleic Acids Research, 31, 4755–4761.

Ferre’, F., Ponty, Y., Lorenz, W.A., and Clote, P. 2007. DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using T&F Cat # C6847 Chapter: 3 page: 67 date: August 5, 2009 nucleotide, dihedral angle and base-pairing similarities. Nucleic Acids Research, 35, W659–W668.

Harrison, A.M., South, D.R., Willett, P., and Artymiuk, P.J. 2003. Representation, Searching and Discovery of Patterns of Bases in Complex RNA Structures, Journal of Computer-Aided Molecular Design, 17(8), 537–549.

Hershkovitz, E., Tannenbaum, E., Howerton, S.B., Sheth, A., Tannenbaum, A., and Williams, L.D. 2003. Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. Nucleic Acids. Research, 31(21), 6249–6257.

Hsiao, C., Mohan, S., Hershkovitz, E., Tannenbaum, A., and

Williams, L.D. 2006. Single nucleotide RNA choreography. Nucleic Acids Research, 34(5), 1481-1491.

Huang, H.-C., Nagaswamy, U., and Fox, G.E. 2005. The application of cluster analysis in the intercomparison of loop structures in RNA. RNA, 11, 412- 423.

Klein, D.J., Schmeing, T.M., Moore, P.B., and Steitz, T.A. 2001. The kinkturn: a new RNA secondary structure motif. The EMBO Journal, 20(15), 4214–4221.

Leontis, N.B., Lescoute, A., and Westhof, E. 2006. The building blocks and motifs of RNA architecture. Current Opinion in Structural Biology, 16, 279–287.

Leontis, N.B., and Westhof, E. 1998. A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. Journal of Molecular Biology, 283, 571–583.

Lescoute, A., and Westhof, E. 2006. Topology of three-way junctions in folded RNAs. RNA, 12, 83–93.

Sarver, M., Zirbel , C.L., Stombaugh, J., Mokdad, A., and Leontis, L.B. 2006. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. Journal Math Biology, doi: 10.1007/s00285-007-0110.

Shapiro, B.A., Yingling, Y.G., Kasprzak, W., and Bindewald, E. 2007. Bridging the gap in RNA structure prediction. Current Opinion in Structural Biology, 17(2), 157–165.

Tamura, M. et al. 2004. Scor: Structural classification of RNA. Nucleic Acids Research, 32, 182–184.

Wadley, L.M., and Pyle, A.M. 2004. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. Nucleic Acid Research, 32, 6650-6659.

4 Chapter 4. Protein Structure Classification Using Machine Learning Methods

[1] http://www.cathdb.info/

[2] The UniProt Consortium. The universal protein resource (UniProt). Nucleic Acids Research, 36:D190-195, 2008.

[3] P. Baldi, S. Brunak, Y. Chauvin, C. Anderson, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification:an overview. Bioinformatics, 16(5):412-424, 2000.

[4] HM. Berman, J. Westbrook, and Z. Feng. The protein data bank. Nucleic Acids Research, 28(1):235-242, 2000. T&F Cat # C6847 Chapter: 4 page: 86 date: August 5, 2009

[5] C.H.Q. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17(4):349- 358, 2001.

[6] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.H. Kim. Recognition of a protein fold in the context of the structural classification of proteins (scop) classification. Proteins, 35:401-407, 1999.

[7] Y. Freund, and R.E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119-139, 1997.

[8] J.H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. Annals of Statistics, 28(2):337-407, 2000.

[9] R.A. Friesner. Computational Methods for Protein Folding: A Special Volume of Advances in Chemical Physics. Wiley-IEEE, Hoboken, NJ, 2002.

[10] C. Hadley, and D.T. Jones. A systematic comparison of protein structure classifications: Scop, cath and fssp. Biological Science, 7(9):1099-1112, 1999.

[11] C.D. Huang, C.T. Lin, and N.R. Pal. Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification. IEEE Transactions on NanoBioscience, 2(4):221-232, 2003.

[12] E. Ie, J. Weston, W.S Noble, and C. Leslie.

Multi-class protein fold recognition using adaptive codes.ACM International Conference, 119:329–336, 2005.

[13] National Human Genome Research Institute. http: //www. genome.

[14] D.T. Jones. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. Molecular Biology, 287(4):797–815, 1999.

[15] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. Bioinformatics, 14(10):846–856, 1998.

[16] S. Lee and R.M. Kil. Multilayer feedforward potential function network. International Joint Conference on Neural Networks, 1:161–171, 1988.

[17] C. J. Lin and C.W. Hsu. A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks, 13(2):415–425, 2002.

[18] L.J. McGuffin, and D.T. Jones. Improvement of the genthreader method for genomic fold recognition. Bioinformatics, 19(7):874–881, 2003.

[19] J. Moody, and C.J. Darken. Fast learning in networks of locally tuned processing units. Neural Computing, 1(2):281–294, 1989. T&F Cat # C6847 Chapter: 4 page: 87 date: August 5, 2009

[20] O. Okun. Protein fold recognition with k-local hyperplane distance nearest neighbor algorithm. Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, 47–53, 2004.

[21] H. Rangwala, and G. Karypis. Building multiclass classifiers for remote homology detection and fold recognition. BMC Bioinformatics, 16(7):455, 2006.

[22] B. Rost and C. Sander. Prediction of protein secondary structure at better 70% accuracy. Journal of Molecular Biology, 232(2):584–599, 1993.

[23] V. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[24] I. Witten, and E. Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, Publishers Inc. San Fransisco, CA, USA, 2005.

[25] J. Xu, Y. Xu, K. Dongsup, and M. Li. Raptor: Optimal protein threading by linear programming. Bioinformatics and Computational Biology, 1(1):95-117, 2003.

[26] Y. Xu and D. Xu. Protein threading using prospect: design and evaluation. Proteins: Structure, Function, and Genetics, 40(3):343-354, 2000.

[27] Y. Krishnaraj, and C.K. Reddy. Boosting methods for protein fold recognition: An empirical comparison. BIBM IEEE International Conference on Bioinformatics and Biomedicine, 393-396, 2008.

5 Chapter 5. Protein Surface Representation and Comparison: New Approaches in Structural Proteomics

Arakaki, A.K. and J. Skolnick. Large-scale assessment of the utility of lowresolution protein structures for biochemical function assignment. Bioinformatics, 20, 2004, 1087–1096.

Baldacci, L., M. Golfarelli, A. Lumini, and S. Rizzi. Clustering techniques for protein surfaces. Pattern Recogn., 39, 2006, 2370–2382.

Binkowski, T. Andrew, L. Adamian, and J. Liang. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J. Mol. Biol., 332(2), 2003 505–526.

Bock, M. E., C. Garutti, and C. Guerra. Discovery of similar regions on protein surfaces. J. Comput. Biol., 14(3), 2007, 285–299.

Canterakis, N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. In Proceedings of the 11th Scandinavian Conference on Image Analysis, Kangerlussuaq, Greenland, 1999, 85–93.

Chen, D. Y., M. Ouhyoung, X. P. Tian, Y. T. Shen, and M. Ouhyoung. On visual similarity based 3D model retrieval. In Proceedings of Eurographics 2003. Granada, Spain, 2003, 223–232.

Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic-acids. Science, 221, 1983, 709–713.

Connolly, M. L. Shape complementarity at the hemoglobin alpha I beta I subunit interface. Biopolymers, 25, 1986, 1229–1247.

de Alarc, P. A., A. D. Pascual-Montano, and J. M. Carazo. Spin Images and Neural Networks for Efficient Content-Based Retrieval in 3D Object Databases. In CIVR ’02: Proceedings of the International Conference on Image and Video Retrieval. London, UK: Springer-Verlag, 2002, 225–234.

Elad, M., A. Tal, and S. Ar. Directed search in a 3d objects database using svm. Technical report, HP Laboratories, Israel, 2000.

Ferre`, Fabrizio, G. Ausiello, A. Zanzoni, and M.

Helmer-Citterich. Functional annotation by identification of local surface similarities: a novel tool for structural genomics. BMC Bioinform., 6, 2005, 194–208.

Fischer, D., S. L. Lin, H. L. Wolfson, and R. Nussinov. A geometry-based suite of molecular docking processes. J. Mol. Biol., 248(2), 1995, 459–477. T&F Cat # C6847 Chapter: 5 page: 107 date: August 5, 2009

Funkhouser, T., and P. Shilane. Partial matching of 3D shapes with prioritydriven search. In Proceedings of the Fourth Eurographics Symposium on Geometry Processing. Carligari, Sardinia, Italy. ACM International Conference Proceeding Series. Eurographics Association, Aire-la-Ville, Swizerland, Vol. 256, June 26–28, 2006, 131–142.

Gerstein, M. A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites. Acta Crystallogr., A48, 1992, 271–276.

Halperin, I., B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins, 47, 2002, 409–443.

Hofbauer, C., H. Lohninger, and A. Aszo´di. SURFCOMP: a novel graph-based approach to molecular surface comparison. J. Chem. Inf. Comput. Sci., 44(3), 2004, 837–847.

Johnson, A. E., and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Anal. Mach. Intell., 21(5), 1999, 433–449.

Kinoshita, K., J. Furui, and H. Nakamura. Identification of protein functions from a molecular surface database, eF-site. J. Struct. Funct. Genomics, 2(1), 2002, 9–22.

Kahraman, A., R. J. Morris, R. A. Laskowski, and J. M. Thornton. Shape variation in protein binding pockets and their ligands. J. Mol. Biol., 368(1), 2007, 283–301.

Kazhdan, M., T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In SGP ’03: Proceedings of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. Aire-la-Ville, Switzerland. Eurographics Association, 2003, 156–164.

Koenderink, J. J., and A. J. van Doorn. Surface shape and curvature scales. Image Vision Comput., 10(8), 1992, 557–564.

Laga, H., H. Takahashi, and M. Nakajima. Spherical wavelet descriptors for content-based 3D model retrieval. In SMI ’06: Proceedings of the IEEE International Conference on Shape Modeling and Applications, Matsushima, Japan, 2006, 75–85.

Leifman, G., S. Katz, A. Tal, and R. Meir. Signatures of 3D models for retrieval. In 4th Israel Korea Bi-National Conference on Geometric Modeling and Computer Graphics, Tel-Aviv, Israel, 2003, 159–163.

Lin, S. L., R. Nussinov, D. Fischer, and H. J. Wolfson. Molecular surface representations by sparse critical points. Proteins, 18(1), 1994, 94–101. T&F Cat # C6847 Chapter: 5 page: 108 date: August 5, 2009

Mademlis, A., A. Axenopoulos, P. Daras, D. Tzovaras, and M. G. Strintzis. 3D content-based search based on 3D Krawtchouk moments. In 3DPVT ’06: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT’06). Washington, DC: IEEE Computer Society, 2006, 743–749.

Mak, L., S. Grandison, and R. J Morris. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. J. Mol. Graph. Model., 26(7), 2008, 1035–1045.

Masek, B. B., A. Merchant, and J. B. Matthew. Molecular skins: a new concept for quantitative shape matching of a protein with its small molecule mimics. Proteins, 17(2), 1993, 193–202.

Novotni, M., and R. Klein. 3D Zernike descriptors for content based shape retrieval. In The 8th ACM Symposium on Solid Modeling and Applications, Seattle, Washington, 2003.

Novotni, M., and R. Klein. Shape retrieval using 3D Zernike descriptors. Computer-Aided Design 36, 11, 2004, 1047–1062.

Ohbuchi, R., M. Nakazawa, and T. Takei. Retrieving 3D shapes based on their appearance. In MIR ’03: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval. New York, NY: ACM Press, 2003, 39–45.

Osada, R., T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. ACM Trans. Graph., 21(4), 2002, 807–832.

Pawlowski, K., and A. Godzik. Surface map comparison: studying function diversity of homologous proteins. J. Mol. Biol., 309, 2001, 793–800.

Pickering, S. J., A. J. Bulpitt, N. Efford, N. D. Gold, and D. R. Westhead. AIbased algorithms for protein surface comparisons. Comput. Chem., 26(1), 2001, 79–84.

Poirrette, A. R., P. J. Artymiuk, D. W. Rice, and P. Willett. Comparison of protein surfaces using a genetic algorithm. J. Comput. Aided Mol. Des., 11(6), 1997, 557–569.

Ritchie, D. W., and G. J. L. Kemp. Protein docking using spherical polar Fourier correlations. Proteins, 39, 2000, 178–194.

Rosen, M., S. L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. Protein Eng., 11(4), 1998, 263–277.

Sael, L., D. La, B. Li, R. Rustamov, and D. Kihara. Rapid comparison of properties on protein surface. Proteins, 73, 2008a, 1–10. T&F Cat # C6847 Chapter: 5 page: 109 date: August 5, 2009

Sael, L., B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, and D. Kihara. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins, 72, 2008b, 1259–1273.

Shentu, Z., M. Al Hasan, C. Bystroff, and M.J. Zaki. Context shapes: efficient complementary shape matching for protein-protein docking. Proteins, 70(3), 2008, 1056–1073.

Shindyalov, I. N., and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng., 11(9), 1998, 739–747.

Tangelder, J. W. H., and R. C. Veltkamp. A survey of content based 3D shape retrieval methods. In SMI ’04: Proceedings of the Shape Modeling International 2004 (SMI’04). Washington, DC: IEEE Computer Society, 2004, 145–156.

Wang, X. Alpha-surface and its application to mining protein data. In Proceedings of the 2001 IEEE International Conference on Data Mining, 2001, 659–662.

Yu, M., I. Atmosukarto, W. K. Leow, Z. Huang, and R. Xu. 3D model retrieval with morphing-based geometric and topological feature maps. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, 656–661.

Zhang, C., and T. Chen. Efficient feature extration for 2D/3D objects in mesh representation. In Proceedings of the 2001 International Conference on Image Processing (ICIP 2001). Thessaloniki, Greece, 2001, October 7–10.

6 Chapter 6. Advanced Graph Mining Methods for Protein Analysis

[1] Jacq, B. Protein function from the perspective of molecular interaction and genetic networks. Bioinformatics, 2, 2001, 38–50.

[2] Hunter, L., ed. Artificial Intelligence and Molecular Biology. The MIT Press Classics Series and AAAI Press, Cambridge, Massachusetts, USA, 1993.

[3] Pevsner, J. Bioinformatics and Functional Genomics. Wiley-Liss, Hoboken, NJ, 2003.

[4] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. The protein data bank. Nucleic Acids Research, 28, 2000, 235–242.

[5] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. Scop:a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 247, 1995, 536–540.

[6] Pearl, F., Bennett, C., Bray, J., Harrison, A., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J.M., Orengo, C. The CATH: an extended protein family resource for structural and functional genomics. Nucleic Acids Research, 31, 2003, 452–455. T&F Cat # C6847 Chapter: 6 page: 133 date: August 5, 2009

[7] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., ODonovan, C., Redaschi, N., Yeh., L.S.L. UniProt: The universal protein knowledgebase. Nucleic Acids Research, 32, 2004, D115–D119.

[8] Simons, K.T., Strauss, C., Baker, D. Prospects for ab initio protein structural genomics. Journal of Molecular Biology, 306, 2001, 1191–1199.

[9] Fiser, A., Sali, A. Modeller: generation and refinement of homology models. Methods in Enzymology, 374, 2003, 461–491.

[10] Kelley, L.A., MacCallum, R.M., Sternberg, M.J.E. Enhanced genome annotation using structural profiles in the program 3d-pssm. Journal of Molecular Biology, 299, 2000, 501–522.

[11] Bourne, P.E., Weissig, H., eds. Structural Bioinformatics. Wiley-Liss, Hoboken, NJ, 2003.

[12] Consortium, T.U. The universal protein resource. Nucleic Acids Research, 36, 2008, 190–195.

[13] Bairoch, A., Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Research, 24, 1996, 21- 25.

[14] Holm, L., Sander, C. Mapping the protein universe. Science, 273, 1996, 595–603.

[15] Siddiqui, A.S., Dengler, U., Barton, G.J. 3Dee: a database of protein structural domains. Bioinformatics, 17, 2001, 200–201.

[16] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., et al. New developments in the InterPro database. Nucleic Acids Research, 35, 2007, D224–D228.

[17] Liew, A.W.C., Yan, H., Yang, M. Data mining for bioinformatics. In: Bioinformatic Technologies, Y.P.P. Chen (ed.). Springer-Verlag Berlin Heidelberg Germany, 2005, 63–106.

[18] Zhang, Q., Veretnik, S., Bourne, P.E. Overview of structural bioinformatics. In: Bioinformatic Technologies, Y.P.P. Chen (ed.). Springer-Verlag Berlin Heidelberg Germany, 2005, 15–44.

[19] Nooren, I.M.A., Thornton, J.M. Diversity of protein–protein interaction. EMBO Journal, 22, 2003, 3486–3492.

[20] Yan, X. Mining, indexing and similarity search in large graph data sets. PhD Dissertation, University of Illionis, Urbana-Champaign, IL, 2006. T&F Cat # C6847 Chapter: 6 page: 134 date: August 5, 2009

[21] Cook, D.J., Holder, L.B. Substructure discovery using minimum description length and back-ground knowledge. Journal of Artificial Intelligence Research, 1, 1994, 231–255.

[22] Yoshida, K., Motoda, H., Indurkhya, N. Graph-based induction as a unified learning frame-work. Journal of Applied Intelligence, 4, 1994, 297- 328.

[23] Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In: Proceedings of 20th International Conference on Very Large Data Bases (VLDB'94), Santiago de Chile, Chile, 1994, 487–499.

[24] Cook, S. The complexity of theorem-proving procedures. In: Proceeding of 3rd ACM Symposium on Theory of Computing (STOC'71), New York, NY, USA, 1971, 151–158.

[25] Washio, T., Motoda, H. State of the art of graph-based data mining. SIGKDD Explorations Newsletter, 5, 2003, 59–68.

[26] Holder, L.B., Cook, D.J., Djoko, S. Substructure discovery in the subdue system. In: Proceeding of the AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94), Seattle, WA, 1994, 169–180.

[27] Inokuchi, A., Washio, T., Motoda, H. An a priori-based algorithm for mining frequent substructures from graph data. In: Proceedings of 2000 European Principles and Practice of Knowledge Discovery in Database (PKDD'00), Lyon, France, 2000, 13–23.

[28] Kuramochi, M., Karypis, G. Frequent subgraph discovery. In: Proceedings of 2001 International Conference on Data Mining (ICDM'01), San Jose, CA, 2001, 313–320.

[29] Yan, X., Han, J. gSpan: Graph-based substructure pattern mining. In: Proceedings of 2002 International Conference on Data Mining (ICDM'02), Maebashi, Japan, 2002, 721–724.

[30] Yan, X., Han, J. CloseGraph: mining closed frequent graph patterns. In: Proceedings of 2003 International Conference of Knowledge Discovery and Data Mining (KDD'03), Washington, DC, 2003, 286–295.

[31] Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., Tropsha, A. Mining spatial motifs from protein structure graphs. In: Proceedings of 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB'04), San Diego, California, USA, 2004, 308–315.

[32] Huan, J., Wang, W., Bankyopadhyay, D., Snoeyink, J., Prins, J., Tropsha, A. Mining protein family specific residue packing patterns from protein T&F Cat # C6847 Chapter: 6 page: 135 date: August 5, 2009 structure graphs. In: Proceeding of 8th International Conference of Research

in Computational Molecular Biology (RECOMB’04), San Diego, California, USA, 2004, 308–315.

[33] Nijssen, S., Kok, J.N. A quickstart in frequent structure mining can make a difference. In: Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’04), Seattle, WA, USA, 2004, 647–652.

[34] Chen, C., Yan, X., Zhu, F., Han, J. gApprox: mining frequent approximate patterns from a massive network. In: Proceeding of 7th IEEE International Conference on Data Mining (ICDM’07), Omaha, NE, USA, 2007, 445–450.

[35] Papadopoulos, A.N., Lyritsis, A., Manolopoulos, Y. SkyGraph: an algorithm for important subgraph discovery in relational graphs. Data Mining and Knowledge Discovery, 17, 2008, 57–76.

[36] Li, X.L., Tan, S.H., Foo, C.S., Ng, S.K. Interaction graph mining for protein complexes using local clique merging. Genome Informatics, 16, 2005, 260–269.

[37] Chen, J., Hsu, W., Lee, M.L., Ng, S.K. NeMoFinder: dissecting genomewide protein-protein interactions with meso-scale network motifs. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD06), Philadelphia, PA, 2006, 106– 115.

[38] Koyuturk, M., Kim, Y., Subramaniam, S., Szpankowski, W., Grama, A. Detecting conserved interaction patterns in biological networks. Journal of Computational Biology, 13, 2006, 1299–1322.

[39] Weskamp, N., Hullermeier, E., kuhn, K., Klebe, G. Multiple graph alignment for the structural analysis of protein active sites. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4, 2007, 310–320.

[40] McGarry, K., Chambers, J., Oatley, G. A multi-layered approach to protein data integration for diabetes research. Artificial Intelligence in Medicine, 41, 2007, 129–143.

[41] Wangikar, P.P., Tendulkar, A.V., Ramya, S., Mali, D.N., Sarawagi, S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. Journal of Molecular Biology, 326, 2003, 955–978.

[42] Deng, H., Chen, G., Yang, W., Yang, J.J. Predicting calcium-binding sites in proteins a graph theory and

geometry approach. Proteins: Structure, Function, and Genetics, 64, 2006, 34–42. T&F Cat # C6847 Chapter: 6 page: 136 date: August 5, 2009

[43] Canutescu, A.A., Shelenkov, A.A., Roland L. Dunbrack, J. A graphtheory algorithm for rapid protein side-chain prediction. Protein Science, 12, 2003, 2001–2014.

[44] Wuchty, S. Scale-free behavior in protein domain networks. Molecular Biology and Evolution, 18, 2001, 1694–1702.

[45] Ye, Y., Godzik, A. Comparative analysis of protein domain organization. Genome Research, 14, 2004, 343–353.

[46] Wagner, A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes.Molecular Biology and Evolution, 18, 2001, 1283–1292.

[47] Koyuturk, M., Grama, A., Szpankowski, W. An efficient algorithm for detecting frequent subgraphs in biological networks. Bioinformatics, 20, 2004, i200–i207.

[48] Bernstein, F. C., Koetzle. T. F., Williams, G. J., Meyer, E. F. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. Journal of Molecular Biology, 112, 1977, 535–542.

7 Chapter 7. Predicting Local Structure and Function of Proteins

[1] S. Ahmad, M. Michael Gromiha, and A. Sarai. Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics, 20(4):477–486, 2004.

[2] S. Ahmad and A. Sarai. Pssm-based prediction of DNA binding sites in proteins. BMC Bioinformatics, 6:33, 2005.

[3] R. Ahmed, H. Rangwala, and G. Karypis. Toptmh: Topology predictor for transmembrane alpha-helices. In European Conference in Machine Learning, R. Goebel, J. Siekmann, and W. Wahlster (Eds.). Antwerp, T&F Cat # C6847 Chapter: 7 page: 158 date: August 5, 2009 Belgium, 2008, Proceedings of European Conference in Machine Learning, Antwerp, Belgium, 2008, 23–28.

[4] S. F. Altschul, L. T. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research, 25(17):3389– 402, 1997.

[5] J. Cheng, M. J. Sweredoski, and P. Baldi. Accurate prediction of protein disordered regions by mining protein structure data. Data Mining and Knowledge Discovery, 11(3):213–222, 2005.

[6] G. E. Crooks, J. Wolfe, and S. E. Brenner. Measurements of protein sequence-structure correlations. Proteins: Structure, Function, and Genetics, 57:804–810, 2004.

[7] A. G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins, 41(3):271–287, 2000.

[8] Z. Dosztnyi, V. Csizmok, P. Tompa P, and I. Simon. Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics, 21(16):3433–3434, 2005.

[9] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic. Intrinsic disorder and protein function. Biochemistry, 41(21):6573–6582, 2002.

[10] C. Etchebest, C. Benros, S. Hazout, and A. de Brevern. A structural alphabet for local protein structures: improved prediction methods. Proteins: Structure, Function, and Bioinformatics, 59:810–827, 2005.

[11] S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi. Poodle-l: a two-level svm prediction system for reliably predicting long disordered regions. Bioinformatics, 23(16):2046–2053, 2007.

[12] T. Joachims. Advances in kernel methods: support vector learning. In Making Large-scale SVM Learning Practical, Joachims (ed.). MIT Press, Cambridge, 1999.

[13] D. T Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics, 23(5):538– 544, 2007.

[14] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22:2577–2637, 1983.

[15] R. Karchin, M. Cline, Y. Mandel-Gutfreund, and K. Karplus. Hidden markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins, 51(4):504–514, 2003.

[16] G. Karypis. Yasspp: better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins, 64(3):575–586, 2006.

[17] C. Kauffman, H. Rangwala, and G. Karypis. Improving homology models for protein-ligand binding sites. In LSS Comput Syst Bioinformatics Conference, number 08-012, San Francisco, CA, 2008. Proceedings of LSS Comput Syst Bioinformatics Conference, number 08–012, San Francisco, CA, 2008.

[18] A. Kernytsky and B. Rost. Static benchmarking of membrane helix predictions. Nucleic Acids Research, 31(13):3642–3644, 2003.

[19] A. R. Kinjo, K. Horimoto, and K. Nishikawa. Predicting absolute contact numbers of native protein structure from amino acid sequence. Proteins: Structure, Function, and Bioinformatics, 58(1):158–165, 2005.

[20] A. R. Kinjo and K. Nishikawa. Crnpred: highly accurate prediction of onedimensional protein structures by large-scale critical random networks. BMC Bioinformatics, 7(401), 2006.

[21] D. Lee, O. Redfern, and C. Orengo. Predicting protein

function from sequence and structure. Nature Reviews Molecular Cell Biology, 8(12):995- 1005, 2007.

[22] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 247:536–540, 1995.

[23] M. N. Nguyen and J. C. Rajapakse. Two-stage support vector machines to protein relative solvent accessibility prediction. In Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, David, W. Corne, Jagath C. Rajapakse, and L. Gwenn Volkert (Eds.). 67–72, 2004.

[24] Y. Ofran, V. Mysore, and B. Rost. Prediction of dna-binding residues from sequence. Bioinformatics, 23(13):347–353, 2007.

[25] T. Ohlson and A. Elofsson. Profnet, a method to derive profile-profile alignment scoring functions that improves the alignments of distantly related proteins. BMC Bioinformatics, 6(253), 2005.

[26] G. Pollastri, P. Baldi, P. Farselli, and R. Casadio. Prediction of coordination number and relative solvent accessibility in proteins. Proteins: Structure, Function, and Genetics, 47:142–153, 2002. T&F Cat # C6847 Chapter: 7 page: 160 date: August 5, 2009

[27] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural network and profiles. Proteins: Structure, Function, and Bioinformatics, 47:228–235, 2002.

[28] H. Rangwala and G. Karypis. frmsdpred: Predicting local rmsd between structural fragments using sequence information. Proteins, 72(3):1005- 1018, 2008.

[29] H. Rangwala, C. Kauffman, and G. Karypis. A generalized framework for protein sequence annotation. In Proceedings of the NIPS Workshop on Machine Learning in Computational Biology, G. Chechik, C. Leslie, W. Noble, Gunnar Ra¨tsch, Quaid Morris, K. Tsuda (Eds.). Vancouver, Canada, 2007.

[30] B. Rost. Phd: predicting 1d protein structure by profile based neural networks. Methods in Enzymology, 266:525–539, 1996.

[31] J. Song and K. Burrage. Predicting residue-wise contact orders in proteins by support vector regression. BMC Bioinformatics, 7(425), 2006.

[32] H. Tjong and Huan-Xiang Zhou. Displar: an accurate method for predicting dna-binding sites on protein surfaces. Nucleic Acids Research, 35(5): 1465–1477, 2007.

[33] V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins: Structure, Function, and Genetics, 41(3):415–427, 2000.

[34] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer Verlag, New Jersey, 1995.

[1] E. Arner, M. Tammi, A.-N. Tran, E. Kindlund, and B. Andersson. DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. BMC Bioinformatics, 7(1):155, 2006.

[2] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. Science, 297(5583):1003–1007, 2002.

[3] D. Bartels, S. Kespohl, S. Albaum, T. Druke, A. Goesmann, J. Herold, O. Kaiser et al. BACCardI—a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. Bioinformatics, 21(7):853–859, 2005.

[4] L. Breiman. Random forests. Machine Learning, 45:5–32, 2001.

[5] J.-H. Choi, S. Kim, H. Tang, J. Andrews, D. G. Gilbert, and J. K. Colbourne. Machine learning approach to combined evidence validation of genome assemblies. Bioinformatics, 24(6):744–750, 2008.

[6] I. M. Dew, B. Walenz, and G. Sutton. A tool for analyzing mate pairs in assemblies (TAMPA). Journal of Computational Biology, 12(5):497–513, 2005.

[7] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning, 40(2):139–157, 2000.

[8] R. O. Duda, and P. E. Hart. Bayes decision theory. In Pattern Classification and Scene Analysis. John Wiley, New York, NY, 10–43, 1973.

[9] M. L. Engle, and C. Burks. GenFrag 2.1: new features for more robust fragment assembly benchmarks. Computer Application in the Biosciences, 10(5):567–568, 1994.

[10] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult et al. The minimal gene complement of Mycoplasma genitalium. Science, 270(5235):397–404, 1995.

[11] D. G. Gilbert. DroSpeGe: rapid access database for new

Drosophila species genomes. Nucleic Acids Research, 35(suppl1):D480-485, 2007.

[12] D. Gordon, C. Abajian, and P. Green. Consed: a graphical tool for sequence finishing. Genome Research, 8(3):195-202, 1998.

[13] P. Green. Phrap. unpublished. http://www.phrap.org.

[14] D. Heckerman, D. Geiger, and D. Chickering. Learning bayesian networks: the combination of knowledge and statistical data. Machine Learning, 20:197-243, 1995.

[15] J. Kececioglu, and J. Ju. Separating repeats in dna sequence assembly. In RECOMB ’01: Proceedings of the Fifth Annual International Conference on Computational Biology. ACM, New York, NY, 176-183, 2001.

[16] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen et al. Mapping and sequencing of structural variation from eight human genomes. Nature, 453:56-64, 2008.

[17] S. Kim, H. Tang, and E. R. Mardis. Advances in Genome Sequencing Technology and Algorithms. Artech House, Norwood, MA, 2007.

[18] S. Kim, and Y. Kim. A fast multiple string pattern matching algorithm. In Proceedings of 17th AoM/IAoM Conference on Computer Science, San Diego, CA, 44-49, 1999.

[19] S. Kim, L. Liao, M. P. Perry, S. Zhang, and J.-F. Tomb. A computational approach to sequence assembly validation. unpublished, 2001. http://bio.informatics.indiana.edu/sunkim/papers/sav.ps.

[20] S. Kim, L. Liao, and J.-F. Tomb. A probabilistic approach to sequence assembly validation. In Zaki, M.J., Toivoven, H., and Wang, J.T. (Eds.) Proceedings of the ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD’01), San Francisco, CA, 38-43, 2001.

[21] S. Kim, and A. M. Segre. AMASS: A structured pattern matching approach to shotgun sequence assembly. Journal of Computational Biology, 6(2):163-186, 1999.

[22] E. W. Myers. Towards simplifying and accurately formulating fragment assembly. Journal of Computational Biology, 2(2):275-290, 1995.

[23] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, et al. A whole-genome assembly of Drosophila. Science, 287(5461):2196–2204, 2000.

[24] A. Phillippy, M. Schatz, and M. Pop. Genome assembly forensics: finding the elusive mis-assembly. Genome Biology, 9(3):R55, 2008.

[25] M. Pop, S. L. Salzberg, and M. Shumway. Genome sequence assembly: algorithms and issues. Computer, 35(7):47–54, 2002.

[26] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.

[27] A. Samad, E. F. Huff, W. Cai, and D. C. Schwartz. Optical mapping: a novel, single-molecule approach to genomic analysis. Genome Research, 5(1):1–4, 1995. T&F Cat # C6847 Chapter: 8 page: 185 date: August 5, 2009

[28] M. P. Samanta, W. Tongprasit, and V. Stolc. In-depth query of large genomes using tiling arrays. Methods Molecular Biology, 377:163–174, 2007.

[29] M. Schatz, A. Phillippy, B. Shneiderman, and S. Salzberg. Hawkeye: an interactive visual analytics tool for genome assemblies. Genome Biology, 8(3):R34, 2007.

[30] M. T. Tammi, E. Arner, T. Britton, and B. Andersson. Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs. Bioinformatics, 18(3):379–388, 2002.

[31] H. Tang. Genome assembly, rearrangement and repeats. Chemistry Reviews, 107(8):3391–3406, 2007.

[32] I. H. Witten, and F. Eibe. Data Mining. Hanser Fachbuch, 2001.

[33] D. Zhi, U. Keich, P. Pevzner, S. Heber, and H. Tang. Correcting baseassignment errors in repeat regions of shotgun assembly. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(01):54- 64, 2007.

[34] A. V. Zimin, D. R. Smith, G. Sutton, and J. A. Yorke. Assembly reconciliation. Bioinformatics, 24(1):42–45, 2008.

9 Chapter 9. Mining Patterns of Epistasis in Human Genetics

Andrew, A.S., Nelson, H.H., Kelsey, K.T., Moore, J.H., Meng, A.C., Casella, D.P., Tosteson, T.D., Schned, A.R., Karagas, M.R. 2006. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking, and bladder cancer susceptibility. Carcinogenesis 27, 1030–37. T&F Cat # C6847 Chapter: 9 page: 201 date: August 5, 2009

Andrew, A.S., Karagas, M.R., Nelson, H.H., Guarrera, S., Polidoro, S., Gamberini, S., Sacerdote, C., Moore, J.H., Kelsey, K.T., Demidenko, E., Vineis, P., Matullo, G. 2008. DNA repair polymorphisms modify bladder cancer risk: a multi-factor analytic strategy. Human Heredity 65, 105–18.

Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D. 1998. Genetic Programming: An Introduction : On the Automatic Evolution of Computer Programs and Its Applications. San Francisco, CA: Morgan Kaufmann Publishers.

Bateson, W. 1909. Mendel’s Principles of Heredity. Cambridge, UK: Cambridge University Press.

Breiman, L. 2001. Random Forests. Machine Learning 45, 5–32.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., et al. 2005. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology 28, 171–82.

Cook, N.R., Zee, R.Y., Ridker, P.M. 2004. Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. Statisics in Medicine 23, 1439–53.

Fisher, R.A. 1918. The correlations between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433.

Fogel, G.B., Corne, D.W. 2003. Evolutionary Computation in Bioinformatics. San Francisco, CA: Morgan Kaufmann Publishers.

Freitas, A. 2001. Understanding the crucial role of attribute interactions. Artificial Intelligence Review 16, 177–99.

Freitas, A. 2002. Data Mining and Knowledge Discovery with Evolutionary Algorithms. New York, NY: Springer.

Hahn, L.W., Ritchie, M.D., Moore, J.H. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19, 376–82.

Jakulin, A., Bratko, I. 2003. Analyzing attribute interactions. Lecture Notes in Artificial Intelligence 2838, 229–40.

Kira, K., Rendell, L.A. 1992. A practical approach to feature selection. In: Sleeman, D.H., Edwards, P. (Eds.) Proceedings of the Ninth International Workshop on Machine Learning (pp. 249–256). San Francisco, CA: Morgan Kaufmann Publishers.

Kononenko, I. 1994. Estimating attributes: analysis and extension of Relief. Proceedings of the European Conference on Machine Learning (pp. 171- 182). New York, NY: Springer.
T&F Cat # C6847 Chapter: 9 page: 202 date: August 5, 2009

Koza, J.R. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA: The MIT Press.

Lewontin, R.C. 1974. The analysis of variance and the analysis of causes. American Journal of Human Genetics 26, 400–11.

Lou, X.Y., Chen, G.B., Yan, L., Ma, J.Z., Zhu, J., Elston, R.C., Li, M.D. 2007. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. American Journal of Human Genetics 80(6), 1125–37.

Lunetta, K.L., Hayward, L.B., Segal, J., Van Eerdewegh, P. 2004. Screening large-scale association study data: exploiting interactions using random forests. BMC Genetetics 5, 32.

McGill, W.J. 1954. Multivariate information transmission. Psychometrica 19, 97–116.

McKinney, B.A., Reif, D.M., Ritchie, M.D., Moore, J.H. 2006. Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics 5(2), 77–88.

Michalski, R.S. 1983. A theory and methodology of inductive learning. Artificial Intelligence 20, 111–61.

Millstein, J., Conti, D.V., Gilliland, F.D., Gauderman, W.J. 2006. A testing framework for identifying susceptibility genes in the presence of epistasis. American Journal of Human Genetics 78(1), 15–27.

Mitchell, T.M. 1997. Machine Learning. Boston, MA: McGraw-Hill.

Moore, J.H. 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity 56, 73–82.

Moore, J.H. 2004. Computational analysis of gene-gene interactions in common human diseases using multifactor dimensionality reduction. Expert Review of Molecular Diagnostics 4, 795–803.

Moore, J.H. 2007. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: Zhu, X., Davidson, I. (Eds.), Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data (pp. 17–30). Hershey, PA: IGI Global.

Moore, J.H., Gilbert, J.C., Tsai, C.-T., Chiang, F.T., Holden, W., Barney, N., White, B.C. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of Theoretical Biology 241, 252–61. T&F Cat # C6847 Chapter: 9 page: 203 date: August 5, 2009

Moore, J.H., Ritchie, M.D. 2004. The challenges of whole-genome approaches to common diseases. Journal of the American Medical Association 291, 1642–43.

Moore, J.H., White, B.C. 2007a. Tuning ReliefF for genome-wide genetic analysis. Lecture Notes in Computer Science 4447, 166–75.

Moore, J.H., White, B.C. 2007b. Genome-wide genetic analysis using genetic programming. The critical need for expert knowledge. In: Rick R., Terence S., Bill W., (Eds.). Genetic Programming Theory and Practice IV (pp. 11-28). New York, NY: Springer.

Moore, J.H., Williams, S.W. 2002. New strategies for identifying gene-gene interactions in hypertension. Annals of Medicine 34, 88–95.

Moore, J.H., Williams, S.W. 2005. Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. BioEssays 27, 637–46.

Pattin, K.A., White, B.C., Barney, N., Gui, J., Nelson, H.H., Kelsey, K.T., Andrew, A.S., Karagas, M.R., Moore, J.H. 2008. A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. Genetic Epidemiology 33(1), 87–94.

Phillips, P.C. 1998. The language of gene interaction. Genetics 149, 1167–71.

Pierce, J.R. 1980. An Introduction to Information Theory: Symbols, Signals, and Noise. New York, NY: Dover.

Reif, D.M., Motsinger, A.A., McKinney, B.A., Crowe Jr, J., Moore, J.H. 2006. Feature selection using random forests for the integrated analysis of multiple data types. Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (pp. 171–178). New York, NY: IEEE Press.

Ritchie, M.D., Hahn, L.W., Moore, J.H. 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, phenocopy, and genetic heterogeneity. Genetic Epidemiology 24, 150–57.

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H. 2001. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. American Journal of Human Genetics 69, 138–47.

Robnik-S˘ikonja, M., Kononenko, I. 2001. Comprehensible interpretation of Relief’s estimates. In: Carla E. Brodley and Andrea Pohoreckyj D. (Eds.) Proceedings of the Eighteenth International Conference on Machine Learning (pp. 433–440). San Francisco, CA: Morgan Kaufmann Publishers. T&F Cat # C6847 Chapter: 9 page: 204 date: August 5, 2009

Robnik-S˘iknja, M., Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53, 23–69.

Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H. 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology 31, 306–15.

Waddington, C.H. 1942. Canalization of development and the inheritance of acquired characters. Nature 150, 563–65.

White, B.C., Gilbert, J.C., Reif, D.M., Moore, J.H. 2005. A statistical comparison of grammatical evolution strategies in the domain of human genetics. Proceedings of the IEEE Congress on Evolutionary Computing (pp. 676– 682). New York, NY: IEEE Press.

10 Chapter 10. Discovery of Regulatory Mechanisms from Gene Expression Variation by eQTL Analysis

Aten JE, Fuller TF, Lusis AJ, Horvath S. 2008. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Syst Biol 2: 34.

Ball RD. 2001. Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. Genetics 159(3): 1351–1364.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57(1): 289–300.

Bing N, Hoeschele I. 2005. Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics 170(2): 533–542.

Breitling R, Amtmann A, Herzyk P. 2004. Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinform 5: 34.

Brem RB, Kruglyak L. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci USA 102(5): 1572–1577.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296(5568): 752–755.

Brem RB, Storey JD, Whittle J, Kruglyak L. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436(7051): 701–703.

Broman KW, Speed TP. 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J R Stat Soc Ser B 64: 641–656.

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT et al. 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. Nat Genet 37(3): 225–232.

Chen Y, Zhu J, Lum PY, Yang X, Pinto S et al. 2008. Variations in DNA elucidate molecular networks that cause disease. Nature 452(7186): 429- 435.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J et al. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37(3): 233–242. T&F Cat # C6847 Chapter: 10 page: 225 date: August 5, 2009

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M et al. 2005. Mapping determinants of human gene expression by regional and genomewide association. Nature 437(7063): 1365–1369.

Churchill GA, Doerge RW. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138(3): 963–971.

DeCook R, Lall S, Nettleton D, Howell SH. 2006. Genetic regulation of gene expression during shoot development in Arabidopsis. Genetics 172(2): 1155- 1164.

Doerge RW, Churchill GA. 1996. Permutation tests for multiple loci affecting a quantitative character. Genetics 142(1): 285–294.

Doss S, Schadt EE, Drake TA, Lusis AJ. 2005. Cis-acting expression quantitative trait loci in mice. Genome Res 15(5): 681–691.

Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. Science 303(5659): 799–805.

Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C et al. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet 2(8): 1182–1192.

Jansen RC, Stam P. 1994. High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136(4): 1447–1455.

Jansen RC, Nap JP. 2001. Genetical genomics: the added value from segregation. Trends Genet 17(7): 388–391.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G et al. 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 29(4): 389–395.

Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD. 2006. Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics 62(1): 19–27.

Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G et al. 2007. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci USA 104(5): 1708–1713.

Kline RB. 2004. Principles and Practice of Structural Equation Modeling. New York, NY: The Guilford Press.

Klose J, Nock C, Herrmann M, Stuhler K, Marcus K et al. 2002. Genetic analysis of the mouse brain proteome. Nat Genet 30(4): 385–393.

Kruglyak L. 2008. The road to genome-wide association studies.Nat Rev Genet 9(4): 314–318. T&F Cat # C6847 Chapter: 10 page: 226 date: August 5, 2009

Kulp DC, Jagalur M. 2006. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. BMC Genomics 7: 125.

Lander ES, Botstein D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121(1): 185–199.

Lee SI, Pe’er D, Dudley AM, Church GM, Koller D. 2006. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc Natl Acad Sci USA 103(38): 14062–14067.

Li H, Lu L, Manly KF, Chesler EJ, Bao L et al. 2005. Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet 14(9): 1119–1125.

Li H, Chen H, Bao L, Manly KF, Chesler EJ et al. 2006. Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. Hum Mol Genet 15(3): 481–492.

Liu B, de la Fuente A, Hoeschele I. 2008. Gene network inference via structural equation modeling in genetical genomics experiments. Genetics 178(3): 1763–1776.

Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P et al. 2004. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75(6): 1094–1105.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG et al. 2004. Genetic analysis of genome-wide variation in human gene expression. Nature 430(7001): 743–747.

Petretto E, Mangion J, Pravanec M, Hubner N, Aitman TJ. 2006a. Integrated gene expression profiling and linkage analysis in the rat. Mamm Genome 17(6): 480–489.

Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK et al. 2006b. Heritability and tissue specificity of expression quantitative trait loci. PLoS Genet 2(10): e172.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. 2002. Hierarchical organization of modularity in metabolic networks. Science 297(5586): 1551–1555.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. Nat Rev Genet 7(11): 862–872.

Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ. 2008. Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. Plant Cell 20(5): 1199–1216. T&F Cat # C6847 Chapter: 10 page: 227 date: August 5, 2009

Sampson JN, Self SG. 2008. Identifying trait clusters by linkage profiles: application in genetical genomics. Bioinformatics 24(7): 958–964.

Schadt EE. 2005. Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. Curr Opin Biotechnol 16(6): 647–654.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422(6929): 297–302.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S et al. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37(7): 710–717.

Segal E, Shapira M, Regev A, Pe’er D, Botstein D et al. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34(2): 166–176.

Sieberts SK, Schadt EE. 2007. Inferring causal associations between genes and disease via the mapping of expression quantitative trait loci. In: Balding DJ, Bishop M, Cannings

C, editors. Handbook of Statistical Genetics. John Wiley & Sons, Ltd, West Sussex, England, 296–326.

Stein CM, Song Y, Elston RC, Jun G, Tiwari HK et al. 2003. Structural equation model-based genome scan for the metabolic syndrome. BMC Genet 4 Suppl 1: S99.

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 100: 9440–9445.

Storey JD, Akey JM, Kruglyak L. 2005. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3(8): e267.

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S et al. 2005. Genome-wide associations of gene expression variation in humans. PLoS Genet 1(6): 0695–0704.

Sun W, Yu T, Li KC. 2007. Detection of eQTL modules mediated by activity levels of transcription factors. Bioinformatics 23(17): 2290–2297.

Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T. 2008. eQED: an efficient method for interpreting eQTL associations using protein networks. Mol Syst Biol 4: 162.

Thaller G, Hoeschele I. 2000. Fine-mapping of quantitative trait loci in half-sib families using current recombinations. Genet Res 76(1): 87–104.

Tu Z, Wang L, Arbeitman MN, Chen T, Sun F. 2006. An integrative approach for causal gene identification and gene regulatory pathway inference. Bioinformatics 22(14): e489–e496. T&F Cat # C6847 Chapter: 10 page: 228 date: August 5, 2009

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y et al. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4(10): 1–15.

Visscher PM, Thompson R, Haley CS. 1996. Confidence intervals in QTL mapping by bootstrapping. Genetics 143(2): 1013–1020.

Wang S, Zheng T, Wang YJ. 2007. Transcription activity hot spot, is it real or an artifact? BMC Proc, 1 Suppl 1: S94.

Wu C, Delano DL, Mitro N, Su Sv, Janes J, et al. 2008. Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. PLoS Genet 4(5): e1000070.

Yu T, Li KC. 2005. Inference of transcriptional regulatory network by twostage constrained space factor analysis. Bioinformatics 21(21): 4033–4038.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E et al. 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35(1): 57–64.

Zeng ZB. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc Natl Acad Sci USA 90(23): 10972– 10976.

Zeng ZB, Kao CH, Basten CJ. 1999. Estimating the genetic architecture of quantitative traits. Genet Res 74(3): 279–289.

Zhu J, Zhang B, Smith EN, Drees B, Brem RB et al.. 2008. Integrating largescale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40(7): 854–861.

Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW et al.. 2004. An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res 105(2-4): 363–374.

11 Chapter 11. Statistical Approaches to Gene Expression Microarray Data Preprocessing

Affymetrix. 2007. Statistical Algorithms Reference Guide, Data Analysis Fundamentals Technical Manual. Affymetrix, Inc., Santa Clara, CA.

Affymetrix. 2005. Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Affymetrix, Inc., Santa Clara, CA.

Affymetrix. 2003. Technical note: design and performance of the GeneChip Human Genome U133 plus 2.0 and Human Genome U133A plus 2.0 arrays. Affymetrix, Inc., Santa Clara, CA.

Affymetrix. 2002. GeneChip Expression Analysis: Data Analysis Fundamentals. Affymetrix, Inc., Santa Clara, CA.

Affymetrix. 1996. Microarray Analysis Suite Version 4 User Guide, Affymetrix, Inc., Santa Clara, CA.

A˚strand M. 2003. Contrast normalization of oligonucleotide arrays. Journal of Computational Biology 10(1):95–102.

Bjork K, Kafadar K. 2007. Order dependence in expression values, variance, detection calls and differential expression in Affymetrix GeneChips. Bioinformatics 23(21):2873–2880.

Bolstad BM. 2004. Low level analysis of high-density oligonucleotide array data: Background, normalization and summarization [dissertation]. University of California at Berkeley.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193.

Calza S, Valentini D, Pawitan Y. 2008. Normalization of oligonucleotide arrays based on the least-variant set of genes. BMC Bioinformatics 9:140: doi:10.1186/1471-2105-9-140.

Chen Z, McGee M, Liu Q, Scheuermann RH. 2007. A distribution free summarization method for Affymetrix GeneChip arrays. Bioinformatics 23(3):321–327. T&F Cat # C6847 Chapter: 11 page: 253 date: August 5, 2009

Chen Z, McGee M, Liu Q, Kong M, Deng Y, Scheuermann RH.

2009. A distribution free convolution method for background correction of oligonucleotide microarray data. BMC Genomics (in press).

Cope LM, Irizarry RA, Jaffee H, Wu Z, Speed TP. 2004. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 1(1):1–13.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Research 33(20):e175.

Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavare´ S, Ritchie ME. 2008. BMC Bioinformatics 9:85: doi:10.1186/1471-2105-9-85.

Dudoit S, Yang YH, Callow MJ, and Speed TP. 2002. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Statistica Sinica 12(1):111–139.

Gentleman RC, Carey VJ, Bates DM, Bolstad BM, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Lacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5:R80, doi:10.1186/gb-2004-5-10-r80.

Gentleman RC, Carey VJ, Huber W, Irizarry RA, Dudoit S (eds). 2005. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York, NY: Springer.

Harbig J, Sprinkle R, Enkemann SA. 2005. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. Nucleic Acids Research 18:33(3):e31.

Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. 2001. Maximumlikelihood estimation of optimal scaling factors for expression array normalization. Proc. SPIE 4266, 132–141.

Hochreiter S, Clevert DA, Obermayer K. 2006. A new summarization method for affymetrix probe level data. Bioinformatics 22(8):943–949.

Hoffmann R, Seidl T, Dugas M. 2002. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biology 3(7): doi:10.1186/gb-2002-3-7-research0033.

Huber W, Von Heydebreck A, Su¨ltmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18(Suppl. 1):S96– S104. T&F Cat # C6847 Chapter: 11 page: 254 date: August 5, 2009

Irizarray RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003a. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 31:e15.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249–264.

Kong YM, Chen Z, Cai J, Scheuermann R. 2007. Use of gene ontology as a tool for assessment of analytical algorithms with real data sets: Impact of revised Affymetrix CDF annotation. 7th International Workshop on Data Mining. Bioinformatics, 64–72.

Kong YM, Chen Z, Qian Y, McClellan E, McGee M, Scheuermann RH. 2009. Objective selection of the optimal microarray analysis pipeline. In preparation.

Lee JA, Sinkovits RS, Mock D, Rab EL, Cai J, Yang P, Saunders B, Hsueh RC, Choi S, Subramaniam S, Scheuermann RH. In collaboration with the Alliance for Cellular Signaling. 2006. Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. BMC Bioinformatics 7:237.

Li C, Wong HW. 2001a. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proceedings of the National Academy of Sciences. 98:31–36.

Li C, Wong HW. 2001b. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology 2: doi:10.1186/gb-2001-2-8-research0032.

McGee M, Chen Z. 2006a. New Spiked-In Probe Sets for the Affymetrix HGU-133A Latin Square Experiment. COBRA Preprint Series. Article 5.

http://biostats.bepress.com/cobra/ps/art5.

McGee M, Chen Z. 2006b. Parameter estimation for the exponential-normal convolution model for background correction of Affymetrix GeneChip data. Statistical Applications in Genetics and Molecular Biology 5(1): Article 24. DOI:10.2202/1544-6115.1237.

Mosteller F, Tukey J. 1977. Data Analysis and Regression. Reading, MA: Addison-Wesley.

Quackenbush J. 2002. Microarray data normalization and transformation. Nature Genetics 32:496–501.

Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel, H. 2000. Normalization strategies for cDNA microarrays. Nucleic Acids Research 28(10):e47. T&F Cat # C6847 Chapter: 11 page: 255 date: August 5, 2009

Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, et al. 2006. The MicroArray Quality Control (MAQC) project shows interand intraplatform reproducibility of gene expression measurements. Nature Biotechnology 24(9):1151–1161.

Therneau TM, Ballman KV. 2008. What does PLIER really do? Cancer Informatics: 6, 423–431.

Warren P, Taylor D, Martini PGV, Jackson J, Bienkouska J. 2007. PANP—a new method of gene detection on oligonucleotide expression arrays. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, 108–115.

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. 2004. A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 99:909–917.

Wu Z, Irizarry RA. 2005. A statistical framework for the analysis of microarray probe-level data. Johns Hopkins University, Department of Biostatistics Working Papers. Working Paper 73 http://www.bepress. com/jhubiostat/paper73.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research 30(4):e15; doi:10.1093/nar/30.4.e15.

12 Chapter 12. Application of Feature Selection and Classification to Computational Molecular Biology

[1] Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403(6769):503–511, 2000. T&F Cat # C6847 Chapter: 12 page: 287 date: August 5, 2009

[2] Allwein E., Schapire R., and Singer Y. Reducing multiclass to binary. Journal of Machine Learning Research, 1:113–141, 2000.

[3] Almuallim H., and Dietterich T.G. Learning with many irrelevant features. In Proceedings of the 9 th National Conference on Artificial Intelligence. MIT Press, Cambridge, MA, 1991.

[4] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., and Levine A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 96(12):6745–6750, 1999.

[5] Bafna V., Halldo´rsson B.V., Schwartz R., Clark A.G., and Istrail S. Haplotypes and informative SNP selection algorithms: Don’t block out information. In Proceedings of the 7 th Annual International Conference on Research in Computational Molecular Biology. RECOMB’03, Berlin, Germany, 19–27, 2003.

[6] Ben-Dor A., Bruhn L., Friedman N., Nachman I., Schummer M., and Yakhini Z. Tissue classification with gene expression profiles. Journal of Computational Biology, 7(3–4):559–83, 2000.

[7] Bertolazzi P., Felici G., Festa P., and Lancia G. Logic classification and feature selection for biomedical data Computer and Mathematics with Applications, 55(5):889–899,2008.

[8] Bertolazzi P., and Felici, G. Learning to classify species with barcodes. IASI Technical Report, 665, 2007.

[9] Bertolazzi P., Felici G., and Festa P. Logic based methods for SNPs tagging and reconstruction. Computer and Operation Research, revision in process, 2007.

[10] Blaxter M., Mann J., Chapman T., Thomas F., Whitton

C., Floyd R., and Abebe E. Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society B, 360(1462):1935- 1943, 2005.

[11] Blaxter M. Molecular systematics: counting angels with DNA. Nature 421:122–124, 2003.

[12] Blaxter M. The promise of a molecular taxonomy. Philosophical Transactions of the Royal Society B, 359:669–679, 2004.

[13] Blaxter M., and Floyd R. Molecular taxonomics forbiodiversity surveys: already a reality, Trends Ecology Evolution, 18:268–269, 2003.

[14] Boros E., Ibaraki T., and Makino K. Logical analysis of binary data with missing bits. Artificial Intelligence, 107:219–263, 1999. T&F Cat # C6847 Chapter: 12 page: 288 date: August 5, 2009

[15] Brown B., Emberson R.M., and Paterson A.M. Mitochondrial COI and II provide useful markers for Weiseana (Lepidoptera, Hepialidae) species identification. Bulletin of Entomological Research, 89:04, 287–293, 1999.

[16] Brown M., Grundy W.N., Lin D., Christianini N., Sugnet C.W., Furey T.S., Ares M. (Jr.), and Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, 97(1):262–267, 2000.

[17] Bucklin A., Guarnieri M., Hill R.S.,Bentley A.M., and Kaartvedt S. Taxonomic and systematic assessment of planktonic copepods using mitochondrial COI sequence variation and competitive, species-specific PCR. Hydrobiology, 401:239–254, 1999.

[18] Chang C-J., Huang Y-T., and Chao K-M. A greedier approach for finding tag SNPs. Bioinformatics, 22(6):685–691, 2006.

[19] Charikar M., Guruswami V., Kumar R., Rajagopalan S., and Sahai A. Combinatorial feature selection problems. In Proceedings of the 41st Annual Symbosium on FOCS 2000, IEEE Computer Society, Washington, DC, USA, 631.

[20] Cristianini N., and Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK, 2000.

[21] Dasarathy B.V. (ed). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society, Los Alamitos, CA, 1991.

[22] DasGupta B., Konwar K.M., Mandoiu I.I., and Shvartsman A.A. Highly scalable algorithms for robust string barcoding. International Conference on Computational Science 2:1020–1028, 2005.

[23] Dash M., and Liu H. Feature selection for classification. Intelligent Data Analysis, I(3):131–156, 1997.

[24] Dietterich T.G., and Bakiri G. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research, 2:263–286, 1995.

[25] Felici G., de Angelis V., and Mancinelli G. Feature selection for data mining. In Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques, G. Felici and E. Triantaphyllou (eds). Springer, New York, USA, 227–252, 2006.

[26] Felici G., and Truemper K. A minsat approach for learning in logic domains. INFORMS Journal on Computing, 13(3):1–17, 2001. T&F Cat # C6847 Chapter: 12 page: 289 date: August 5, 2009

[27] Felici G., and Truemper K. The Lsquare system for mining logic data. In Encyclopedia of Data Warehousing and Mining, J. Wang (ed.), vol. 2. Idea Group Inc., Hershey PA, USA, 693–697, 2006.

[28] Freund Y., and Schapire R.E. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119–139, 1997.

[29] Furey T.S., Christianini N., Duffy N., Bednarski D.W., Schummer M., and Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10):906–914, 2000.

[30] Garey M.R., and Johnson D.S. Computer and Intractability: A Guide to the Theory of NP-Completeness. Freeman, San Francisco, CA, 1979.

[31] Gennari J.H., Langley P., and Fisher D. Models of

incremental concept formation. Artificial Intelligence, 40:11-61, 1989.

[32] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439):531- 537, 1999.

[33] Gordon G.J., Jensen R.V., Hsiao Li-Li, Gullans S.R., Blumenstock J.E., Ramaswamy S., Richards W.G., Sugarbaker D.J., and Bueno R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research, 62:4963-4967, 2002.

[34] Rash, S., and Gusfield, D. String barcoding: Uncovering optimal virus signatures. In Proceedings 6 th Annual International Conference on Computational Biology, Washington, DC, USA, 254-261, 2002.

[35] Guyon I., Weston J., Barnhill S., and Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389-422, 2002.

[36] Hall M.A. Correlation-based feature selection for machine learning. In Proceedings of the 17 th International Conference on Machine Learning. Morgan Kaufmann, CA, 2000.

[37] Halldrsson B.V., Istrail S., and De La Vega F.M. Optimal selection of SNP markers for disease association studies. Human Heredity, 58:190- 202, 2004.

[38] Halperin E., Kimme G., and Shamir R. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. Bioinformatics, 21:195- 203, 2005. T&F Cat # C6847 Chapter: 12 page: 290 date: August 5, 2009

[39] Hajibabaei M., Singer G.A.C., Clare E.L., Paul D.N., and Hebert P.D.N. Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. BMC Biology, 5(24):1-7, 2007.

[40] Hebert P.D.N., Cywinska A., Ball S.L., and deWaard J.R. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B, 270:313-321, 2003.

[41] Hebert P.D.N., Penton E.H, Burns J.M, Janzen D.H., and Hallwachs W. Ten species in one: DNA barcoding reveals

cryptic species in the Neotropical skipper butterfly Astraptes fulgerator. Proceedings of the National Academy Sciences USA, 101:14812-14817, 2004.

[42] He J., and Zelikovsky A. Tag SNP selection based on multivariate linear regression, International Conference on Computational Science (2). Lecture Notes in Computer Science 3992:750-757, 2006.

[43] Jia Min X. and Hickey D.A. Assessing the effect of varying sequence length on DNA barcoding of fungi. Molecular Ecology Notes 1, 7(3):365- 373, 2007.

[44] Hu H., Li J., Plank A., Wang H., and Daggard G. A comparative study of classification methods for microarray data analysis. In Proceedings of the 5 th Australasian Conference on Data Mining and Analystics, Sydney, Australia 61:33-37, 2006.

[45] Jirapech-Umpai T., and Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6:148, 2005.

[46] Koller D., and Sahami M. Hierachically classifying documents using very few words. In Machine Learning: Proceedings of the 14 th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 170-178, 1997.

[47] Kuksa P., and Pavlovic V. Kernel methods for DNA barcoding. In Snowbird Learning Workshop, San Juan, Puerto Rico, March 19-22, 2007.

[48] Langley P. Selection of relevant features in machine learning, Artificial Intelligence, 97, 245-271, 1997.

[49] Chengliang Z., Li T., and Mitsunori O. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15):2429-2437, 2004.

[50] Liu H., and Setiono R. A probabilistic approach to feature selection: A filter solution. In Proceedings of the 13 th International Conference on Machine Learning, Bari Italy, July 27, 1996. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 319-327, 1996. T&F Cat # C6847 Chapter: 12 page: 291 date: August 5, 2009

[51] Maniatis N., Collins A., Xu C.F., Mcfhy L.C., Hewett

D.R., Tapper W., Ennis S., Ke X., Morton N.E. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proceedings of the National Academy of Sciences USA, 99:2228-2233, 2002.

[52] Min X.J., and Hickey D.A. DNA barcodes provide a quick preview of mitochondrial genome composition. PLoS ONE, 2(3):e325, 2007.

[53] Montgomery D., and Undem B.L. Drug discovery. CombiMatrix' customizable DNA microarrays. Genetic Engineering News, 22(7):32-33, 2002.

[54] Nanney, D.L. Genes and phenes in Tetrahymena. Bioscience, 32:783-740, 1982.

[55] Oliveira A.L., and Vincetelli A.S., Constructive induction using a nongreedy strategy for feature selection. In Proceedings of the 9 th International Conference on Machine Learning. Morgan Kaufmann, Aberdeen, Scotland, 355-360, 1992.

[56] Pace N.R. A molecular view of microbial diversity and the biosphere. Science, 276:734-740, 1997.

[57] Pasaniuc B., Kentros S., and Mandoiu I.I. Boosting assignment accuracy by combining distanceand character-based classifiers. In The DNA Barcode Data Analysis Initiative: Developing Tools for a New Generation of Biodiversity Data. Paris, France, July 6-8, 2006, unpublished presentation, http://dna.engr.uconn.edu/?page id=21&year=2006.

[58] Paschou P., Mahoney M.W., Javed A., Kidd J.R., Pakstis A.J., Gu S., Kidd K.K. and Drineas P. Intraand interpopulation genotype reconstruction from tagging SNPs. Genome Research, 17:96-107, 2007.

[59] Patil N. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science, 294:1719-1723, 2001.

[60] Petricoin E.F., Ardekani A.M., Hitt B.A., Levine P.J., Fusaro V.A., Steinberg S.M., Mills G.B., Simone C., Fishman D.A., Kohn E.C., and Liotta L.A. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 359(9306):572-577, 2002.

[61] Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA, 1993.

[62] Saccone C., DeCarla G., Gissi C., Pesole G., and Reynes A. Evolutionary genomics in the Metazoa: the mitochondrial DNA as a model system. Gene, 238:195–210, 1999. T&F Cat # C6847 Chapter: 12 page: 292 date: August 5, 2009

[63] Saitou N., and Nei M. The Neighbour-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology Evolution, 4(4):406– 425, 1987.

[64] Schlimmer J.C. Efficiently inducing determinations: a complete and systematic search algorithm that uses optimal pruning. In Proceedings of the 10 th International Conference on Machine Learning. Morgan Kaufmann, Amherst, MA, 284–290, 1993.

[65] Schummer M., Ng W.V., Bumgarner R.E., Nelson P.S., Schummer B., Bednarski D.W., Hassell L., Baldwin R.L., Karlan B.Y., and Hood L. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene, 238(2):375–385, 1999.

[66] Sheinvald J., Dom B., and Niblack W. Unsupervised image segmentation using the minimum description length principle. In Proceedings of the 11 th IAPR International Conference on Pattern Recognition, 2:709–712, 1992.

[67] Singh D., Febbo P.G., Ross K., Jackson D.G., Manola J., Ladd C., Tamayo P. et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2):203–209, 2002.

[68] Smith M. A., Woodley N. E., Janzen D. H., Hallwachs W., and Hebert P.D.N. DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proceedings of the National Academy of Sciences, 103:3657–3662, 2006.

[69] Vapnik V.N. Statistical Learning Theory. Wiley, New York, NY, 1998.

[70] Veer L.J., Dai H., van de Vijver M.J., He Y.D., Hart A.A., Mao M., Peterse H.L. et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871):530–536, 2002.

[71] Xiong H., and Chen X. Kernel-based distance metric

learning for microarray data classification. BMC Bioinformatics, 7:299, 2006.

[72] Ye J., Li T., Xiong T., and Janardan R. Using uncorrelated discriminant analysis for tissue classification with gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(4):181- 190, 2004.

[73] Zhang K., Qin Z.S., Liu J.S., Chen T., Waterman M.S., and Sun F. Haplotype block partitioning and Tag SNP selection using genotype data and their applications to association studies. Genome Research, 14:908- 916, 2004.

[74] Zhang K., Qin Z.S., Liu J.S., Chen T., Waterman M.S., and Sun F. HapBlock: haplotype block partitioning and Tag SNP selection software using a set of dynamic programming algorithms. Bioinformatics, 21:131- 134, 2005.

[75] Orsenigo C., and Vercellis C. Discrete support vector decision trees via tabu-search. Journal of Computational Statistics and Data Analysis, 47:311–322, 2004.

[76] Orsenigo C., and Vercellis C. Accurately learning from few examples with a polyhedral classifier. Computational Optimization and Applications, 38:235–247, 2007.

[77] Orsenigo C. Gene selection and cancer microarray data classification via mixed-integer optimization. In: Evolutionary computation, machine learning, data mining in bioinformatics 6 th European Conference, EvoBIO 2008, Naples, Italy, 2008. Lecture Notes in Computer Science, 4973:141-152, 2008.

[78] Triantaphyllou E., The OCAT approach for data mining and knowledge discovery. Working Paper, IMSE Department, Louisiana State University, Baton Rouge, LA, 70803-6409, 2001.

13 Chapter 13. Statistical Indices for Computational and Data Driven Class Discovery in Microarray Data

[1] Supplementary material web site. http://www.math.unipa.it/~raffaele/ suppMaterial/chapterDM/. T&F Cat # C6847 Chapter: 13 page: 333 date: August 5, 2009

[2] Validation Work Bench: Valworkbench web page. http://www.math. unipa.it/ ~raffaele/valworkbench/.

[3] J. N. Breckenridge. Replicating cluster analysis: Method, consistency, and validity. Multivariate Behavioral Research, 24(2):147-161, 1989.

[4] S. Datta and S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics, 19:459-466, 2003.

[5] V. Di Gesu´, R. Giancarlo, G. Lo Bosco, A. Raimondi, and D. Scaturro. Genclust: A genetic algorithm for clustering gene expression data. BMC Bioinformatics, 6:289, 2005.

[6] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology, 3, 2002.

[7] S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering solution. Bioinformatics, 19:1090-1099, 2003.

[8] E. B. Fowlkes and C.L. Mallows. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, 78:553-584, 1983.

[9] A. D. Gordon. Null models in cluster validation. InW. Gaul and D. Pfeifer (Eds.), From Data to Knowledge: Theoretical and Practical Aspects of Classification. Springer Verlag, Berlin, 32-44, 1996.

[10] A. D. Gordon. Clustering algorithms and cluster validation. In P. Dirschedl and R. Ostermann, editors. Computational Statistics. PhysicaVerlag, Heidelberg, Germany, 497-512, 1994.

[11] J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. Bioinformatics, 21(15):3201-3212, 2005.

[12] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir. An algorithm for clustering of cDNAs for gene expression analysis using short oligonucleotide fingerprints. Genomics, 66:249-256, 2000.

[13] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, Berlin, 2003.

[14] M. J. L. De Hoon, S. Imoto, and S. Miyano. The C Clustering Library for cDNA Microarray Data. Laboratory of DNA Information Analysis Human Genome Center, Institute of Medical Science, University of Tokyo, 2007.

[15] L. Hubert and P. Arabie. Comparing partitions. Journal of Classification, 2:193–218, 1985. T&F Cat # C6847 Chapter: 13 page: 334 date: August 5, 2009

[16] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. PrenticeHall, Englewood Cliffs, NJ, 1988.

[17] L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, NY, 1990.

[18] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proceedings of the National Academy of Sciences USA, 98:8961–8965, 2001.

[19] W. Krzanowski and Y. Lai. A criterion for determining the number of groups in a dataset using sum of squares clustering. Biometrics, 44:23–34, 1985.

[20] M-Y. Leung, G. M. Marsch, and T. P. Speed. Over and underrepresentation of short DNA words in Herphesvirus genomes. Journal of Computational Biology, 3:345–360, 1996.

[21] F. H. C. Marriot. Practical problems in a method of cluster analysis. Biometrics, 27:501–514, 1971.

[22] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50:159– 179, 1985.

[23] G. W. Milligan and M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research, 21:441–458, 1986.

[24] S. Monti, P. Tamayo, J. Mesirov, and T. Golub.

Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning, 52:91–118, 2003.

[25] I. Priness, O. Maimon, and I. Ben-Gal. Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinformatics, 8:111, 2007.

[26] W. M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66:846–850, 1971.

[27] C. Van Rijsbergen. Information Retrieval, second edition. Butterworths, London, UK, 1979.

[28] R. Shamir and R. Sharan. Algorithmic approaches to clustering gene expression data. In T. Jiang, T. Smith, Y. Xu, and M. Q. Zhang, editors. Current Topics in Computational Biology. MIT Press, Cambridge, MA, 120–161, 2003. T&F Cat # C6847 Chapter: 13 page: 335 date: August 5, 2009

[29] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell, 9:3273–3297, 1998.

[30] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistics. Journal Royal Statistical Society B, 2:411–423, 2001.

[31] X. Wen, S. Fuhrman, G. S. Michaels, G. S. Carr, D. B. Smith, J. L. Barker, and R. Somogyi. Large scale temporal gene expression mapping of central nervous system development. Proceedings of the National Academy of Science USA, 95:334–339, 1998.

[32] M. Yan and K. Ye. Determining the number of clusters with the weighted gap statistics. Biometrics, 63:1031–1037, 2007.

[33] K. Y. Yeung. Cluster analysis of gene expression data. PhD Thesis, University of Washington, WA, 2001.

[34] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo. Validating clustering for gene expression data. Bioinformatics, 17:309–318, 2001.

14 Chapter 14. Computational Approaches to Peptide Retention Time Prediction for Proteomics

Anderson, N. L. and Anderson, N. G. 2002. The human plasma proteome: history, character, and diagnostics prospects. Mol. Cell. Proteomics, 1.11, 845–867.

Baczek, T., Wiczling, P., Marszatt, M., Heyden, Y. V., and Kaliszan, R. 2005. Prediction of peptide at different HPLC conditions from multiple linear regression models. J. Proteome Res., 4, 555–563.

Chabanet, C., and Yvon, M. 1992. Prediction of peptide retention time in reversed-hpase high-performance liquid chromatography. J. Chormatogr. A 599, 211–225.

Gilar, M., Jaworski, A., Olivova, P., and Gebler, J. C. 2007. Peptide retention time prediction applied to proteomic data analysis. Rapid Commun. Mass Spectrom., 21, 2813–2821.

Guo, D., Mant, C. T., Taneja, A. K., Parker, J. M. R., Hodges, R. S. 1986. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides. J. Chromatogr., A359, 499–518.

Krokhin, O. V., Craig, R., Spicer, V., Ens, W., Standing, K. G., Beavis, R. C., and Wilkins, J. A. 2004. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phse HPLC. Mol. Cell. Proteomics, 3, 908–919.

Krokhin O. V. 2006a. Sequence-sepcific retention calculator. Agorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300and 100-A˚ pore size C18 sorbents. Anal. Chem. 78, 7785–7795. T&F Cat # C6847 Chapter: 14 page: 348 date: August 5, 2009

Krokhin, O. V., Ying, S., Cortens, J. P., Ghosh, D., Spicer, V., Ens, W., Standing, K. G., Beavis, R. C., and Wilkins, J. A. 2006b. Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLCMALDI MS/MS. Anal. Chem., 78, 6265–6269.

Ladiwala, A., Xia, F., Luo, Q., Breneman, C. M., and Cramer, S. M. 2006. Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models. Biotechnol. Bioeng., 93, 836–850.

Meek, J. L. 1980. Prediction retention time in high-pressure liquid chromatography on the basis of amino acid composition. Proc. Natl. Acad. Sci. USA. 77, 1632–1636.

Spicer, V., Yamchuk, A., Cortens, J., Sousa, S., Ens, W., Standing, K. G., Wilkins, J. A., and Krokhin, O. V. 2007. Sequence-specific retention calculator. A family of peptide retention time prediction algorithms in reversedphase HPLC: applicability to various chromatographic conditions and columns. Anal. Chem. 79, 8762–8768.

Oh, C., Z˙ak, S. H., Mirzaei, H., Buck, C., Regnier, F. E., and Zhang, X. 2007. Neural network prediction of peptide separation in strong anion exchange chromatography. Bioinformatics, 23, 114–118.

Palmblad, M., Ramstrom, M., Markides, K. E., Hakansson, P., and Bergquist, J. 2002. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. Anal. Chem., 74, 5826–5830.

Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A., Pasˇa-Toli, L., Lipton, M. S., Auberry, K. J., Strittmatter, E. F., Shen, Y., Zhao, R., and Smith, R. D. 2003. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analysis. Anal. Chem., 75, 1039–1048.

Petritis, K., Kangas, L. J., Yan, B., Monroe, M. E., Strittmatter, E. F., Qian, W.-J., Adkins, J. N., Moore, R. J., Xu, Y., Lipton, M. S., Camp II, D. G., and Smith, R. D. 2006. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. Anal. Chem. 78, 5026–5039.

Put, R., Daszykowski, M., Baczek, T., and Heyden, Y. V. 2006. Retention prediction of peptides based on uninformative variable elimination by partial least squares. J. Proteome Res., 5, 1618–1625.

Qiu, R., Zhang, X., Regnier, F. E. 2007. A method for the identification of glycoproteins from human serum by a combination of lectin affinity chromatography along with anion exchange and Cu-IMAC selection of tryptic peptides. J. Chromatogr. B., 845, 143–150. T&F Cat # C6847 Chapter:

Shinoda, K., Sugimoto, M., Yachie, N., Sugiyama, N., Masuda, T., Robert, M., Soga, T., and Tomita, M. 2006. Prediction of liquid chromatographic retention times of peptides generated by protease digestion of the Escherichia coli proteome using artificial neural networks. J. Proteome Res., 5, 3312- 3317.

Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennett, K. P., Cramer, S., and Tugcu, N. 2002. Prediction of protein retention times in anionexchange chromatography systems using support vector regression. J. Chem. Inf. Comput. Sci., 42, 1347–1357.

Strittmatter, E. F., Kangas, L. J., Petritis, K., Mottas, H. M., Anderson, G. A., Shen, Y., Jacobs, J. M., Camp II, D. G., and Smith, R. D. 2004. Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. J. Proteome Res., 3, 760–769.

Vapnik, V. N. 1998. Statistical learning theory. Wiley, New York.

Vapnik, V. N. 2000. The nature of statistical learning theory, 2nd ed. SpringerVerlag, New York.

15 Chapter 15. Inferring Protein Functional Linkage Based on Sequence Information and Beyond

Bowers, P.M., S.J. Cokus, D. Eisenberg and T.O. Yeates. 2004. Use of logic relationships to decipher protein network organization. Science, 306, 2246– 2249.

Craig, R and L. Liao. 2007a. Transductive learning with EM algorithm to classify proteins based on phylogenetic profiles. Int. J. Data Mining Bioinformatics, 1, 337–351.

Craig, R. and L. Liao. 2007b. Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. BMC Bioinformatics, 8, 6.

Craig, R.A. and L. Liao. 2007c. Improving protein-protein interaction prediction based on phylogenetic information using least-squares SVM. Ann. New York Acad. Sci., 1115(1), 154–167.

Craig, R.A., K. Malaviya, K. Balasubramanian and L. Liao. 2008. Inferring functional linkage from residue level co-evolution information. The International Conference on Bioinformatics and Computational Biology (BioComp08), Las Vegas, NV.

Cristianini, N. and J. Shawe-Taylor. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK.

Durbin, R., S.R. Eddy, A. Krogh and G. Mitchison. 1999. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.

Friedrich, T., B. Pils, T. Dandekar, J. Schltz and T. Muller. 2006. Modeling interaction sites in protein domains with interaction profiles hidden Markov models. Bioinformatics, 22, 2851–2857.

Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt and M. Gerstein. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302, 449–453.

Kahsay, R., G. Gao and L. Liao. 2005. Discriminating transmembrane proteins from signal peptides using

SVM-Fisher approach. Proceedings of 4 th International Conference on Machine Learning and Applications (ICMLA), Los Angeles, CA, 151–155.

Kim, Y., M. Koyuturk, U. Topkara, A. Grama and S. Subramaniam. 2006. Inferring functional information from domain co-evolution. Bioinformatics, 22, 40–49. T&F Cat # C6847 Chapter: 15 page: 376 date: August 5, 2009

Liao, L. 2006. Hierarchical profiling, scoring and applications in bioinformatics. In Advanced Data Mining Technologies in Bioinformatics, edited by Hui-Huang Hsu. Idea Group, Inc., Hershey, USA.

Patel, T. and L. Liao. 2007. Predicting protein-protein interaction using Fisher scores extracted from domain profiles. Proceedings of IEEE 7th International Symposium for Bioinformatics and Bioengineering (BIBE), Boston, MA.

Patel, T., M. Pillay, R. Jawa and L. Liao. 2006. Information of binding sites improves prediction of protein-protein interaction. Proceedings of the Fifth International Conference on Machine Learning and Applications (ICMLA’06), Orlando, FL, 205–210.

Pazos, F. M.H. Citterich, G. Ausiello and A. Valencia. 1997. Correlated mutations contain information about protein-protein interaction. J. Mol. Biol., 271(4), 511–523.

Pazos, F. and A. Valencia. 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng., 14, 609–614.

Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA, 96, 4285–4288.

Sato, T., Y. Yamanishi, M. Kanehisa and H. Toh. 2005. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics, 21, 3482–3489.

Scholkopf, B. and A.J. Smola. 2001. Learning with Kernels: Support Vector Machines, Regularizaton, Optimization, and Beyond. The MIT Press, Cambridge, MA.

Scholkopf, B., K. Tsuda and J.P. Vert, editors. 2004.

Kernel Methods in Computational Biology. The MIT Press, Cambridge, MA.

Suykens, J.A.K. and J. Vandewalle. 1999. Least squares support vector machine classifiers. Neural Proc. Lett., 9, 293–300.

Valencia, A and F. Pazos. 2003. Prediction of protein-protein interactions from evolutionary information. In Structural Bioinformatics, edited by P.E. Bourne and H. Weissig. Wiley-Liss, Inc.

Yamanishi, Y., J.-P. Vert and M. Kanehisa. 2004. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics, 20, i363–i370.

16 Chapter 16. Computational Methods for Unraveling Transcriptional Regulatory Networks in Prokaryotes

[1] J. L. Reed, I. Famili, I. Thiele et al. Towards multidimensional genome annotation. Nat Rev Genet, 7(2), 130–41, 2006.

[2] J. J. Faith, B. Hayete, J. T. Thaden et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol, 5(1), 54–66, 2007.

[3] T. S. Gardner, D. di Bernardo, D. Lorenz et al. Inferring genetic networks and identifying compound mode of action via expression profiling. Science, 301(5629), 102–5, 2003.

[4] M. Madan Babu, S. A. Teichmann, and L. Aravind. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol, 358(2), 614–33, 2006.

[5] D. A. Tagle, B. F. Koop, M. Goodman et al. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol, 203(2), 439–55, 1988.

[6] M. J. de Hoon, Y. Makita, K. Nakai et al. Prediction of transcriptional terminators in Bacillus subtilis and related species. PLoS Comput Biol, 1(3), 212–221, 2005.

[7] H. Salgado, G. Moreno-Hagelsieb, T. F. Smith et al. Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci USA, 97(12), 6652–57, 2000.

[8] M. D. Ermolaeva, O. White, and S. L. Salzberg. Prediction of operons in microbial genomes. Nucleic Acids Res, 29(5), 1216–21, 2001.

[9] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. Science, 278(5338), 631–37, 1997.

[10] R. W. Brouwer, O. P. Kuipers, and S. A. van Hijum. The relative value of operon predictions. Brief Bioinform, 9(5), 367–75, 2008.

[11] G. Li, D. Che, and Y. Xu. A universal operon predictor

for prokaryotic genomes. J Bioinform Comput Biol, 7(1), 19–38, 2009.

[12] D. Che, J. Zhao, L. Cai et al. Operon prediction in microbial genomes using decision tree approach. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 135–42, 2007.

[13] I. H. Witten, and E. Frank. Data Mining: Practical machine learning tools and techniques, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

[14] J. R. Quinlan. C4.5 Programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[15] T. Cormen, C. Leiserson, R. Rivest et al. Introduction to algorithms. Cambridge, MA: The MIT Press, 2001.

[16] S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res, 36, Database issue, D120–24, 2008.

[17] G. Z. Hertz, and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15(7–8), 563–77, 1999.

[18] K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides et al. A comparative genomics approach to prediction of new members of regulons. Genome Res, 11(4), 566–84, 2001.

[19] Z. Su, V. Olman, F. Mao et al. Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. Nucleic Acids Res, 33(16), 5156–71, 2005.

[20] L. A. McCue, W. Thompson, C. S. Carmack et al. Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res, 12(10), 1523–32, 2002.

[21] S. Neph, and M. Tompa. MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes. Nucleic Acids Res, 34, Web Server issue, W366–68, 2006. T&F Cat # C6847 Chapter: 16 page: 395 date: August 5, 2009

[22] S. T. Jensen, L. Shen, and J. S. Liu. Combining

phylogenetic motif discovery and motif clustering to predict co-regulated genes. Bioinformatics, 21(20), 3832-39, 2005.

[23] D. Che, G. Li, S. Jensen et al. PFP: a computational framework for phylogenetic footprinting in prokaryotic genomes. Lecture Notes Comput Sci, 4983, 110-21, 2008.

[24] J. van Helden, A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res, 28(8), 1808-18, 2000.

[25] G. Pavesi, P. Mereghetti, G. Mauri et al. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res, 32, Web Server issue, W199-203, 2004.

[26] M. Tompa, N. Li, T. L. Bailey et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol, 23(1), 137-44, 2005.

[27] G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics, 17(Suppl 1), S207- 14, 2001.

[28] K. D. MacIsaac, and E. Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol, 2(4), 201-210, 2006.

[29] X. Liu, D. Brutlag, and J. Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of coexpressed genes. Pac. Symp. Biocomput. 127-138, 2001.

[30] T. L. Bailey, and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2, 28-36, 1994.

[31] C. T. Harbison, D. B. Gordon, T. I. Lee et al. Transcriptional regulatory code of a eukaryotic genome. Nature, 431(7004), 99-104, 2004.

[32] J. Hu, B. Li, and D. Kihara. Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res, 33(15), 4899-913, 2005.

[33] D. Che, S. Jensen, L. Cai et al. BEST: binding-site estimation suite of tools," Bioinformatics, 21(12), 2909-11, 2005.

[34] S. T. Jensen, and J. S. Liu. BioOptimizer: a Bayesian scoring function approach to motif discovery. Bioinformatics, 20(10), 1557–64, 2004.

[35] S. Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res, 24(19), 3836- 45, 1996. T&F Cat # C6847 Chapter: 16 page: 396 date: August 5, 2009

[36] T. Wang, and G. D. Stormo. Combining phylogenetic data with coregulated genes to identify regulatory motifs. Bioinformatics, 19(18), 2369–80, 2003.

[37] A. Sandelin, and W. W. Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol, 338(2), 207–15, 2004.

[38] T. Wang, and G. D. Stormo. Identifying the conserved network of cisregulatory sites of a eukaryotic genome. Proc Natl Acad Sci USA, 102(48), 17400–5, 2005.

[39] Z. S. Qin, L. A. McCue, W. Thompson et al. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. Nat Biotechnol, 21(4), 435–39, 2003.

[40] W. C. Lathe, 3rd, B. Snel, and P. Bork. Gene context conservation of a higher order than operons. Trends Biochem Sci, 25(10), 474–79, 2000.

[41] D. Che, G. Li, F. Mao et al. Detecting uber-operons in prokaryotic genomes. Nucleic Acids Res, 34(8), 2418–27, 2006.

17 Chapter 17. Computational Methods for Analyzing and Modeling Biological Networks

[1] Prz˘ulj N. Graph theory analysis of protein-protein interactions. In Knowledge Discovery in Proteomics. Edited by Jurisica I, Wigle D. CRC Press, Boca Raton, FL, 2005:73–128.

[2] Sharan R, Ulitsky I, Ideker T. Network-based prediction of protein function. Molecular Systems Biology 2007:3(88).

[3] Milenkovic´ T, Prz˘ulj N. Uncovering biological network function via graphlet degree signatures. Cancer Informatics 2008:6:257-273.

[4] Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, Mooney SD. An inte-grated approach to inferring gene-disease associations in humans. Proteins 2008:72(3):1030-1037.

[5] Jonsson P, Bates P. Global topological features of cancer proteins in the human interactome. Bioinformatics 2006:22(18):2291-2297.

[6] Newman MEJ. The structure and function of complex networks. SIAM Review 2003:45(2):167–256.

[7] Milo R, Shen-Orr SS, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science 2002:298:824–827.

[8] Prz˘ulj N, Corneil DG, Jurisica I. Modeling interactome: scale-free or geometric? Bioinformatics 2004:20(18):3508–3515.

[9] Prz˘ulj N. Biological network comparison using graphlet degree distribution. Bioinformatics 2007:23:e177–e183.

[10] Prz˘ulj N, Higham D. Modelling protein-protein interaction networks via a stickiness index. Journal of the Royal Society Interface 2006:3(10):711- 716.

[11] Erdo¨s P, Re´nyi A. On random graphs. Publicationes Mathematicae 1959:6:290–297.

[12] Molloy M, Reed B. A critical point for random graphs with a given degree sequence. Random Structures and Algorithms 1995:6:161–179.

[13] Baraba´si AL, Albert R. Emergence of scaling in random networks. Science 1999:286(5439):509–512.

[14] Watts DJ, Strogatz SH. Collective dynamics of ’small-world’ networks. Nature 1998:393:440–442. T&F Cat # C6847 Chapter: 17 page: 427 date: August 5, 2009 Networks

[15] Penrose M. Geometric Random Graphs. Oxford University Press, USA, 2003.

[16] Milenkovic´ T, Lai J, Prz˘ulj N. GraphCrunch: a tool for large network analyses. BMC Bioinformatics 2008:9(70).

[17] Vazquez A, Flammini A, Maritan A, Vespignani A. Modeling of protein interaction networks. ComPlexUs 2003:1:38–44.

[18] Higham D, Ras˘ajski M, Prz˘ulj N. Fitting a geometric graph to a proteinprotein interaction network. Bioinformatics 2008:24:1093–1099.

[19] Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. Nature Biotechnology 2006:24(4):427–433.

[20] Singh R, Xu J, Berger B. Pairwise global alignment of protein interaction networks by matching neighborhood topology. RECOMB 2007, LNBI 2007:4453:16–31.

[21] Prz˘ulj N, Wigle D, Jurisica I. Functional topology in a network of protein inter-actions. Bioinformatics 2004:20(3):340–348.

[22] King AD, Prz˘ulj N, Jurisica I. Protein complex prediction via cost-based clus-tering. Bioinformatics 2004:20(17):3013–3020.

[23] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, et al. Towards a proteome-scale map of the human proteinprotein interaction network. Nature 2005:437:1173–1178.

18 Chapter 18. Statistical Analysis of Biomolecular Networks

Bader, G.D., and Hogue, C.W. 2003. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2.

Chung, F., and Lu, L. 2006. Complex Graphs and Networks. Providence, RI: American Mathematical Society.

Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102–105.

Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.D., Hao, T., Berriz, G.F., Bertin, N., Huang, J., Chuang, L.S., et al. 2005. Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436, 861–865. T&F Cat # C6847 Chapter: 18 page: 445 date: August 5, 2009

Han, J.D. 2008. Understanding biological functions through molecular networks. Cell Res 18, 224–237.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P., et al. 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430, 88–93.

Han, J.D., Dupuy, D., Bertin, N., Cusick, M.E., and Vidal, M. 2005. Effect of sampling on topology predictions of protein–protein interaction networks. Nat Biotechnol 23, 839–844.

He, H., and Singh, A.K. 2006. GraphRank: statistical modeling and mining of significant subgraphs in the feature space. The 2006 IEEE-WIC-ACM International Conference on Date Mining (ICDM 2006), Hong Kong, Published by IEEE Computer Society Press, 45–59.

Heckerman, D. 1998. A tutorial on learning with Bayesian networks. In: Jordan MI, editor. Learning in Graphical Models. Kluwer Academic, Boston, 301–354.

Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., and Alon, U. 2003. Subgraphs in random networks. Phys Rev E Stat Nonlin Soft Matter Phys 68, 026127.

Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. 2001. Lethality and centrality in protein networks. Nature 411, 41–42.

Koyutu¨rk, M., Szpankowski, W., and Grama, A. 2007. Assessing significance of connectivity and conservation in protein interaction networks. J Comput Biol, 14(6): 747–764.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. 2002. Network motifs: simple building blocks of complex networks. Science 298, 824–827.

Needham, C.J., Bradford, J.R., Bulpitt, A.J., and Westhead, D.R. 2007. A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol 3, e129.

Newman, M.E., and Girvan, M. 2004. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69, 026113.

Papin, J.A., Hunter, T., Palsson, B.O., and Subramaniam, S. 2005. Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6, 99–111.

Pujana, M.A., Han, J.D., Starita, L.M., Stevens, K.N., Tewari, M., Ahn, J.S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., et al. 2007. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet 39, 1338–1349. T&F Cat # C6847 Chapter: 18 page: 446 date: August 5, 2009

Rives, A.W., and Galitski, T. 2003. Modular organization of cellular networks. Proc Natl Acad Sci USA 100, 1128–1133.

Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D.A., and Nolan, G.P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. Science 308, 523–529.

Segre, D., Deluna, A., Church, G.M., and Kishony, R. 2005. Modular epistasis in yeast metabolism. Nat Genet 37, 77–83.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., and Ideker, T. 2005. Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci USA 102, 1974–1979.

Snel, B., Bork, P., and Huynen, M.A. 2002. The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci USA 99, 5890-5895.

Watts, D.J., and Strogatz, S.H. 1998. Collective dynamics of ‘small-world’ networks. Nature 393, 440-442.

Wernicke, S., and Rasche, F. 2006. FANMOD: a tool for fast network motif detection. Bioinformatics 22, 1152-1153.

Wuchty, S., Oltvai, Z.N., and Barabasi, A.L. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 35, 176-179.

Xue, H., Xian, B., Dong, D., Xia, K., Zhu, S., Zhang, Z., Hou, L., Zhang, Q., Zhang, Y., and Han, J.D. 2007. A modular network model of aging. Mol Syst Biol 3, 147.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., and Gerstein, M. 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3, e59.

Yu, H., Zhu, S., Zhou, B., Xue, H., and Han, J.D. 2008. Inferring causal relationships among different histone modifications and gene expression. Genome Res 18, 1314-1324.

Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., and Jarvis, E.D. 2004. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20, 3594-3603.

19 Chapter 19. Beyond Information Retrieval: Literature Mining for Biomedical Knowledge Discovery

Albert, R., and Barabasi, A.-L. 2002. Statistical mechanics of complex networks. Reviews of Modern Physics, 74(1):47–97.

Ando, R. K. 2007. BioCreative II gene mention tagging system at IBM watson. In Proceedings the Second BioCreative Challenge Evaluation Workshop, Madrid, Spain, 101–103.

Aronson, A. R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proc AMIA Symp, Washington, DC, 17–21.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., and Nielsen, H. 2000. Assessing the accuracy of prediction algorithms for classification: An overview. Bioinformatics, 16:412–424.

Bashyam, V., Divita, G., Bennett, D., Browne, A., and Taira, R. 2007. A normalized lexical lookup approach to identifying umls concepts in free text. In Proceedings of the 12th World Congress on Health Informatics MEDINFO (2007), Brisbane, Australia, 545–549.

Bernhardt, P. J., Humphrey, S. M., and Rindflesch, T. C. 2005. Determining prominent subdomains in medicine. In AMIA Annual Symposium Proceeding 2005. American Medical Informatics Association, Washington, DC, 46–50.

Bernstam, E. V., Herskovic, J. R., Aphinyanaphongs, Y., Aliferis, C. F., Sriram, M. G., and Hersh, W. R. 2006. Using citation data to improve retrieval from medline. Journal of the American Medical Informatics Association, 13(1):96–105.

Bhavnani, S., Abraham, A., Demeniuk, C., Gebrekristos, M., Gong, A., Nainwal, S., Vallabha, G., and Richardson, R. 2007. Network analysis of toxic chemicals and symptoms: Implications for designing first-responder systems. In Proceedings of AMIA ’07, Chicago, IL, 51–55.

Bo¨rner, K., Dall’ Asta, L., Ke, W., and Vespignani, A. 2005. Studying the emerging global brain: Analyzing and visualizing the impact of coauthorship teams: Research articles. Complexity, 10(4):57–67.

Camon, E., Barrell, D., Dimmer, E., Lee, V., Magrane, M., Maslen, J., Binns, D., and Apweiler, R. 2005. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics, 6(Suppl 1):S17.

Cantor, M. N., and Lussier, Y. A. 2003. Putting data integration into practice: Using biomedical terminologies to add structure to existing T&F Cat # C6847 Chapter: 19 page: 481 date: August 5, 2009 data sources. In AMIA Annual Symposium Proceedings, Washington, DC, 125–129.

Castelli, S., Meossi, C., Domenici, R., Fontana, F., and Stefani, G. 1993. Magnesium in the prophylaxis of primary headache and other periodic disorders in children. Pediatria Medica e Chirurgica, 15(5):481–488.

Cohen, A. M., and Hersh, W. R. 2005. A survey of current work in biomedical text mining. Brief Bioinformatics, 6(1):57–71.

Cohen, K. B., Acquaah-Mensah, G. K., Dolbey, A. E., and Hunter, L. 2002. Contrast and variability in gene names. In Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. Association for Computational Linguistics, Morristown, NJ, 14–20.

Cutting, D. R., Karger, D., Pedersen, J. O., and Tukey, J. W. 1992. Scatter/ Gather: A cluster-based approach to browsing large document collections. In SIGIR ’92. Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Copenhagen, Denmark, 318–329.

Fang, H. R., Murphy, K., Jin, Y., Kim, J., and White, P. 2006. Human gene name normalization using text matching with automatically extracted synonym dictionaries. In Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language Processing and Biology. Association for Computational Linguistics, Morristown, NJ, 41–48.

Fawcett, T. 2003. Roc graphs: Notes on practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Laboratories, Palo Alto, CA.

Franze´n, K., Eriksson, G., Olsson, F., Asker, L., and Lide´n, P. 2002. Exploiting syntax when detecting protein names in text. In Proceedings of Workshop on Natural Language Processing in Biomedical Applications NLPBA 2002. Nicosia, Cyprus.

Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. 1998.

Toward information extraction: Identifying protein names from biological papers. In Proceedings of the Pacific Symposium on Biocomputing, Hawaii, 705–716.

Gordon, M. D. and Lindsay, R. K. 1996. Toward discovery support systems: a replication, re-examination, and extension of Swanson’s work on literaturebased discovery of a connection between Raynaud’s and fish oil. Journal of the American Society for Information Science, 47(2):116–128.

Gupta, B. M. 1997. Analysis of distribution of the age of citations in theoretical population genetics. Scientometrics, 40(1):139–162. T&F Cat # C6847 Chapter: 19 page: 482 date: August 5, 2009

Hanisch, D., Fluck, J., Mevissen, H., and Zimmer, R. 2003. Playing biology’s name game: Identifying protein names in scientific text. In Proceedings of the Pacific Symposium on Biocomputing, Hawaii, 403–414.

Hatzivassiloglou, V., Duboue´, P. A., and Rzhetsky, A. 2001. Disambiguating proteins, genes, and RNA in text: A machine learning approach. Bioinformatics, 17(Suppl 1):s97–s106.

Hsu, C.-N., Chang, Y.-M., Kuo, C.-J., Lin, Y.-S., Huang, H.-S., and Chung, I.-F. 2008. Integrating high dimensional bi-directional parsing models for gene mention tagging. Bioinformatics, 24(13):i286–i294.

Jensen, L., Saric, J., and Bork, P. 2006. Literature mining for the biologist: From information retrieval to biological discovery. Nature Reviews Genetics, 7:119–129.

Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., and RebholzSchuhmann, D. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics, 9(Suppl 3):S3.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632.

Krauthammer, M., Rzhetsky, A., Morozov, P., and Friedman, C. 2001. Using BLAST for identifying gene and protein names in journal articles. GENE, (259):245–252.

Landauer, T., Foltz, P., and Laham, D. 1988. Introduction to latent semantic analysis. Discourse Processes, 25:259–284.

Lindsay, R. K., and Gordon, M. D. 1999. Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50(7):574–587.

Mane, K., and Bo¨rner, K. 2004. Mapping topics and topic bursts in PNAS. Proceedings of the National Academy of Science, 101:5287–5290.

Mika, S., and Rost, B. 2004. Protein names precisely peeled off free text. Bioinformatics, 20(Suppl 1):i241–i247.

Mostafa, J. 2004. Seeking better web searches. Scientific American, 292(2):51- 71.

Mostafa, J., Mukhopadhyay, S., Lam, W., and Palakal, M. 1997. A multilevel approach to intelligent information filtering: Model, system, and evaluation. ACM Transactions on Information Systems, 368–399.

Newman, M. E. J. 2001. Scientific collaboration networks. i. network construction and fundamental results. Physical Review E, 64(1):016131. T&F Cat # C6847 Chapter: 19 page: 483 date: August 5, 2009

Nicolaisen, J. 2007. Citation analysis. Annual Review of Information Science and Technology, 43:609–641.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1998. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project.

Perez-Iratxeta, C., Wjst, M., Bork, P., and Andrade, M. 2005. G2D: a tool for mining genes associated with disease. BMC Genetics, 6(1):45.

Salton, G., and McGill, M. 1983. Introduction to Modern Information Retrieval, 1st edition. McGraw-Hill, Ohio.

Schuemie, M. J., Weeber, M., Schijvenaars, B. J. A., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B., and Kors, J. A. 2004. Distribution of information in biomedical abstracts and full-text publications. Bioinformatics, 20(16):2597–2604.

Sebastiani, F. 2002. Machine learning in automated text categorization. ACM Computing Survery, 34(1): 1–47.

Seki, K., and Mostafa, J. 2005. A hybrid approach to protein name identification in biomedical texts. Information

Processing and Management, 41(4):723–743.

Seki, K., and Mostafa, J. 2007. Discovering implicit associations between genes and hereditary diseases. The Pacific Symposium on Biocomputing, 12:316– 327.

Seki, K. and Mostafa, J. 2009. Discovering implicit associations among critical biological entities. International Journal of Data Mining and Bioinformatics, 3(2), in press.

Sparck Jones, K. 1972. Statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1):11–20.

Srinivasan, P. 2004. Text mining: generating hypotheses from Medline. Journal of the American Society for Information Science and Technology, 55(5):396–413.

Swanson, D. R. 1986a. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1):7–18.

Swanson, D. R. 1986b. Undiscovered public knowledge. Library Quarterly, 56(2):103–118.

Swanson, D. R. 1987. Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Science, 38(4):228–233. T&F Cat # C6847 Chapter: 19 page: 484 date: August 5, 2009

Swanson, D. R., and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, 91(2):183–203.

Swanson, D. R., Smalheiser, N. R., and Torvik, V. I. 2006. Ranking indirect connections in literature-based discovery: The role of medical subject headings. Journal of the American Society for Information Science and Technology, 57(11): 1427–1439.

Torvik, V. I., Weeber, M., Swanson, D. R., and Smalheiser, N. R. 2005. A probabilistic similarity metric for medline records: A model for author name disambiguation. Journal of the American Society for Information Science and Technology, 56(2):140–158.

Turtle, H., and Croft, W. B. 1991. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, 9(3):187-222.

Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T., Aronson, A. R., and Molema, G. 2003. Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. Journal of the American Medical Informatics Association, 10(3):252–259.

Wilbur, J., Larry, S., and Lorrie, T. 2007. BioCreative 2. Gene mention task. In Proceedings the Second BioCreative Challenge Evaluation Workshop, Madrid, Spain, 7–16.

Yang, Y., and Liu, X. 1999. A re-examination of text categorization methods. In SIGIR ’99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, NY, 42–49.

Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. 2004. Recognizing names in biomedical texts: A machine learning approach. Bioinformatics, 20(7):1178–1190.

20 Chapter 20. Mining Biological Interactions from Biomedical Texts for Efficient Query Answering

[1] Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. and Takeda, K. A text-mining system for knowledge discovery from biomedical documents. IBM Systems Journal, 43(3), 516–533, 2004.

[2] Muller, H. M., Kenny, E. E. and Strenber, P. W. Textpresso: An ontologybased information retrieval and extraction system for biological literature. PloS Biology 2(11), e309, URL: http://www.plosbiology.org, 1984–1998, 2004.

[3] Jensen, L. J., Saric, J. and Bork, P. Literature mining for the biologist: From information retrieval to biological discovery. Nature Reviews Genetics, 7(2), 119–129, 2006.

[4] Allen, J. Natural Language Understanding, 2 nd edition. Pearson Education (Singapore) Pte. Ltd., India, 2004.

[5] Schuler, G.D. et al. Entrez: Molecular biology database and retrieval system. Methods in Enzymology, 266, 141–162, 1996.

[6] Erhardt, R. A-A., Scheider, R. and Blaschke, C. Status of text-mining techniques applied to biomedical text. Drug Discovery Today, 11(7/8), 315–324, 2006.

[7] Fensel, D., Horrocks, I., Harmelen, F. van, McGuinness, D. L. and PatelSchneider, P. OIL: Ontology infrastructure to enable the semantic web. IEEE Intelligent Systems, 16(2), 38–45, 2001. T&F Cat # C6847 Chapter: 20 page: 525 date: August 5, 2009

[8] Bada, M. et al. A short study on the success of the gene ontology. Journal of Web Semantics, 1, 235–240, 2004.

[9] Ashburner, M. et al. Gene ontology: Tool for the unification of biology, the gene ontology consortium. Nature Genetics, 25, 25–29, 2000.

[10] Cohen, A. M. and Hersh, W. R. A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57–71, 2005.

[11] McNaught, J. and Black, W. Information extraction. In Text Mining for Biology and Biomedicine, Ananiadou S. and McNaught J. (eds). Artech House, Norwood, USA. 143–178,

2006.

[12] Riloff, E. and Lehnert, W. Information extraction as a basis for highprecision text classification. ACM Transactions on Information Systems, 12, 296–333, 1994.

[13] Tateisi, Y., Ohta, T., Collier, N., Nobata, C. and Tsujii, J. Building annotated corpus in the molecular-biology domain. In Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, Morgan Kaufmann Publisher, San Francisco, USA, 28–34, 2000.

[14] Gruber, T. R. A translation approach to portable ontology specification. Knowledge Acquisition, 5(2), 199–220, 1993.

[15] Ding, J. et al. Mining medline: Abstracts, sentences or phrases. In Pacific Symposium on Biocomputing, World Scientific, ISBN-10:981024777X, 326–337, 2002.

[16] Sekimizu, T., Park, H. S. and Tsujii, J. Identifying the interaction between genes and genes products based on frequently seen verbs in Medline abstract. Genome Informatics, 9, 62–71, 1998.

[17] Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. Automatic extraction of protein interactions from scientific abstracts. In Pacific Symposium on Biocomputing, World Scientific, ISBN-10:9810241887, 538–549, 2000.

[18] Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T. Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics, 17(2), 155–161, 2001.

[19] Rinaldi, F., Scheider, G., Andronis, C., Persidis, A. and Konstani, O. Mining relations in the GENIA corpus. In Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy, 2004.

[20] Cohen, A. M. and Hersh, W. R. A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1),57–71, 2005. T&F Cat # C6847 Chapter: 20 page: 526 date: August 5, 2009

[21] Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T. Toward information extraction: Identifying protein names from biological papers. In Pacific Symposium on Biocomputing, World Scientific, ISBN-10:9810225784, 707–718,

1998.

[22] Proux, D., Rechenmann, F., Julliard, L., Pillet, V. and Jacq, B. Detecting gene symbols and names in biological texts: A first step toward pertinent information extraction. In Proceedings of the 9th workshop on Genome Informatics, Universal Academy Press, Japan, 72–80, 1998.

[23] Hanisch, D., Fluck, J., Mevissen, H. T. and Zimmer, R. Playing biology's name game: Identifying protein names in scientific text. In Pacific Symposium on Biocomputing, World Scientific, ISBN-10:9812382178, 403–414, 2003.

[24] Rindflesch, T. C., Hunter, L. and Aronson, A. R. Mining molecular binding terminology from biomedical text. In Proceedings of the AMIA Symposium, Washington, American Medical Information Association, USA, 127–131, 1999.

[25] Collier, N., Nobata, C. and Tsujii, J. Extracting the names of genes and gene products with a hidden Markov model. In Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000), Morgan Kaufmann Publisher, San Francisco, USA, 201–207, 2000.

[26] Nobata, C., Collier, N., and Tsujii, J. Automatic term identification and classification in biology texts. In Proceedings of the Natural Language Pacific Rim Symposium, Beijing, China, 369–375, 1999.

[27] Settles, B. ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14), 3191- 3192, 2005.

[28] Schutz, A. and Buitelaar, P. RelExt: A tool for relation extraction from text in ontology extension. In Proceedings of the 4th International Semantic Web Conference (ISWC), Galway. Ireland LNCS-3729, Springer, Berlin, 593–606, 2005.

[29] Yakushiji, A., Teteisi, Y., Miyao, Y. and Tsujii, J. Event extraction from biomedical papers using a full parser. In Pacific Symposium on Biocomputing, World Scientific, ISBN-10:9810245157, 408–419, 2001.

[30] Wagner, C., Cheung, K. S. K. and Rachael, K.F. Building semantic webs for e-government with wiki technology. Electronic Government, 3(1), 36- 55, 2006.

[31] Garc´ıa, R. and Celma, O. Semantic integration and retrieval of multimedia metadata. In Proceedings of the 5th

International Workshop on Knowledge Markup and Semantic Annotation, Galway, Ireland, 69–80, 2005. T&F Cat # C6847 Chapter: 20 page: 527 date: August 5, 2009

[32] Cox, E. A hybrid technology approach to free-form text data mining. URL: http://scianta.com/pubs/AR-PA-007.htm

[33] Castro, A. G., Rocca-Serra, P., Stevens, R., Taylor, C., Nashar, K., Ragan, M. A. and Sansone, S-A. The use of concept maps during knowledge elicitation in ontology development processes—the nutrigenomics use case. BMC Bioinformatics, 7:267, Published online May 25, 2006.

[34] Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J. and Rojas, I. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, Scotland, UK, Professional Book Center, ISBN-0938075934, 659–664, 2005.

[35] Aumann, Y. et al. Circle graphs: New visualization tools for text mining. In Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery. LNCS-1704, Springer-Verlag, UK, 277– 282, 1999.

[36] Abulaish, M. and Dey, L. Biological relation extraction and query answering from Medline abstracts using ontology-based text mining. Data & Knowledge Engineering, Vol. 61(2). Elsevier Science Publishers, Amsterdam, Netherlands, 228–262, 2007.

[37] Jenssen, T-K., Laegreid, A., Komorowski, J. and Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, 28, 21–28, 2001.

[38] Gaizauskas, R., Demetriou, G., Artymiuk, P. J. and Willett, P. Protein structures and information extraction from biological texts: the PASTA system. Bioinformatics, 19(1), 135–143, 2003.

[39] Craven, M. and Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. In Proceedings of the 7 th International Conference on Intelligent Systems for Molecular Biology (ISMB'99), Heidelberg Germany, AAAI Press, ISBN:1-57735-083-9, 77– 86, 1999.

[40] Friedman, C., Kra, P., Yu, H., Krauthammer, M. and Rzhetsky, A. GENIES: A natural-language processing system

for the extraction of molecular pathways from journal
articles. Bioinformatics, 17(Suppl. 1), s74-s82, 2001.

21 Chapter 21. Ontology-Based Knowledge Representation of Experiment Metadata in Biological Data Mining

Aranguren ME, Antezana E, Kuiper M, Stevens R. Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. BMC Bioinform. 2008, 9(Suppl 5):S1.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000, 25(1):25–9.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res. 2007, 35(Database issue):D760–65.

Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008, 67–79.

Bornberg-Bauer E, Paton NW. Conceptual data modeling for bioinformatics. Brief Bioinform. 2002, 3:166–80.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001, 29(4):365–71.

Conenello GM, Zamarin D, Perrone LA, Tumpey T, Palese P. A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence. PLoS Pathog. 2007, 3(10):1414–21.

Cook DL, Wiley JC, Gennari JH. Chalkboard: ontology-based pathway modeling and qualitative inference of disease mechanisms. Pac Symp Biocomput. 2007, 16–27.

Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC et al. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res. 2007, 35(Database issue):D766–70.

Diehl AD, Lee JA, Scheuermann RH, Blake JA. Ontology development for biological systems: immunology. Bioinformatics. 2007, 23(7):913–15.

Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: a technical report. J Am Med Inform Assoc. 2008, 15(2):130–37. T&F Cat # C6847 Chapter: 21 page: 558 date: August 5, 2009

Greene JM, Collins F, Lefkowitz EJ, Roos D, Scheuermann RH, Sobral B, Stevens R, White O, Di Francesco V. National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. Infect Immun. 2007, 75(7):3212–19.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K et al. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004, 32(Database issue): D258–61.

Hucka M, Bolouri H, Finney A, Sauro H, Doyle JHK, Arkin A, Bornstein B et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003, 19:524–31.

Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, Degreef J et al. The functional genomics experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol. 2007, 25(10):1127–33.

Lee JA, Sinkovits RS, Mock D, Rab EL, Cai J, Yang P, Saunders B, Hsueh RC, Choi S, Subramaniam S, Scheuermann RH. Alliance for cellular signaling. Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. BMC Bioinform. 2006, 7:237.

Lee JA, Spidlen J, Atwater S, Boyce K, Cai J, Crosbie N, Dalphin M et al. MIFlowCyt: the minimum information about a flow cytometry experiment. Cytometry: Part A. 2008, 73(10):926–30.

Le Nove`re N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ et al. Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol. 2005, 23(12):1509–15.

Lloyd C, Halstead M, Nielsen P. CellML: its future, present and past. Prog Biophy Mol Biol. 2004, 85:433–50.

Luciano J. PAX of mind for pathway researchers. Drug

Discovery Today. 2005, 10(13):937–42.

Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH. Modular organization of protein interaction networks. Bioinformatics. 2007, 23(2): 207–14.

O’Connor MJ, Shankar RD, Parrish DB, Das AK. Knowledge-data integration for temporal reasoning in a clinical trial system. Int J Med Inform. 2009, 78:577–85. T&F Cat # C6847 Chapter: 21 page: 559 date: August 5, 2009

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007, 35(Database issue):D747–50.

Qian Y, Tchuvatkina D., Spidlen J, Wilkinson P, Garparetto M, Jones AR, Manion FJ, Scheurmann RH, Sekaly RP, and Brinkmann RR. FuGEFlow: data model and markup language for flow cytometry. BMC Bioinform. 2009, submitted.

Ramakrishnan R, Gehrke J. Database Management Systems, 3 rd Edition. McGraw-Hill Co., New York, 2003.

Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform. 2008, 9(1):75–90.

Smith B, Ceusters W, Klagges B, Ko¨hler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biol. 2005, 6(5):R46.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007, 25(11):1251–55.

Squires B, Macken C, Garcia-Sastre A, Godbole S, Noronha J, Hunt V, Chang R, Larsen CN, Klem E, Biersack K, Scheuermann RH. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. Nucleic Acids Res. 2008, 36(Database issue): D497–503.

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol. 2008, 26(8):889–96.

22 Chapter 22. Redescription Mining and Applications in Bioinformatics

[1] R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD' 93), Washington, DC, 207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Databases (VLDB' 94), Santiago de Chile, Chile, 487–499, 1994.

[3] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. Identification of dene regulatory networks by strategic gene disruptions and gene overexpressions. In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, 695–702, 1998.

[4] C.A. Ball, I.A. Awad, J. Demeter, J. Gollub, J.M. Hebert, T. HernandezBoussard, H. Jin, J.C. Matese, M. Nitzberg, F. Wymore, Z.K. Zachariah, P.O. Brown, and G. Sherlock. The Stanford microarray database accomodates additional microarray platforms and data formats. Nucleic Acids Research, 1(33):D580–D582, 2005.

[5] B.-J. Breitkreutz, C. Stark, and M. Tyers. The GRID: The general repository for interaction datasets. Genome Biology, 3(12), 2002.

[6] N.H. Bshouty. Exact learning Boolean functions via the monotone theory. Information and Computation, 123(1):146–153, 1995.

[7] A.E. Carpenter and D.M. Sabatini. Systematic genome-wide screens of gene function. Nature Reviews Genetics, 5(1):11–22, 2004.

[8] S.Y. Chan and D.R. Appling. Regulation of S-adenosylmethionine levels in Saccharomyces cerevisiae. Journal of Biological Chemistry, 278(44):43051–43059, 2003.

[9] U. de Lichtenberg, L.J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. Science, 307(5710):724–727, 2005.

[10] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide

expression patterns. PNAS, 95(25):14863- 14868, 1998.

[11] D.H. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2(2):139–172, 1987.

[12] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell, 11:4241–4257, 2000.

[13] A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, et al. Functional organization of the yeast proteome by systematic analysis of protein Complexes. Nature, 415(6868):141–147, 2002.

[14] K.C. Gunsalus and F. Piano. RNAi as a tool to study cell biology: Building the genome-phenome bridge. Current Opinion in Cell Biology, 17(1):3–8, 2005.

[15] M.A. Matzke and J.A. Birchler. RNAi-mediated pathways in the nucleus. Nature Reviews Genetics, 6(1):24–35, 2005.

[16] M.A. Matzke and A.J.M. Matzke. Planting the seeds of a new paradigm. PLoS Biology, 2(5):582–586, 2004.

[17] R.S. Michalski. Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data into conjunctive concepts. International Journal of Policy Analysis and Information Systems, 4:219–243, 1980.

[18] M. Mizunuma, K. Miyamura, D. Hirata, H. Yokoyama, and T. Miyakawa. Involvement of S-adenosylmethionine in G1 cell cycle regulation in Saccharomyces cerevisiae. PNAS, 101(16):6086–6091, 2004.

[19] L. Parida and N. Ramakrishnan. Redescription mining: Structure theory and algorithms. In Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI’05), Pittsburgh, PA, 837–844, 2005.

[20] Y. Pilpel, P. Sudarsanam, and G.M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. Nature Genetics, 29(2):153–159, 2001.

[21] K.Q. Pu and A.O. Mendelzon. Concise descriptions of subsets of structured sets. ACM Transactions on Database Systems, 30(1):211–248, 2005. T&F Cat # C6847 Chapter: 22

[22] J.R. Quinlan. Induction of decision trees. Machine Learning, 1(1):81–106, 1986.

[23] N. Ramakrishnan, M. Antoniotti, and B. Mishra. Reconstructing formal Temporal models of cellular events using the GO process ontology. In Proceedings of the Eighth Annual Bio-Ontologies Meeting (ISMB’05 Satellite Workshop), Detroit, MI, 2005.

[24] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R.F. Helm. Turning CARTwheels: An alternating algorithm for mining redescriptions. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’04), Seattle, WA, 266–275, 2004.

[25] E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing Conditional activity of expression modules in cancer. Nature Genetics, 36(10):1090–1098, 2004.

[26] E Segal, M. Shapira, A. Regev, D. Pe’er, D. Botstein, D. Koller, and N. Friedman. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics, 34(2):166–176, 2003.

[27] J. Singh, D. Kumar, N. Ramakrishnan, V. Singhal, J. Jervis, A. Desantis, J. Garst, S. Slaughter, M. Potts, and R.F. Helm. Transcriptional response of Saccharomyces cerevisiae to desiccation and rehydration. Applied and Environmental Microbiology, 71(12):8752–8763, 2005.

[28] A. Sturn, J. Quackenbush, and Z. Trajanoski. Genesis: Cluster analysis of microarray data. Bioinformatics, 18(1):207–208, 2002.

[29] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 102(43):15545– 15550, 2005.

[30] N. Sudarsan, J.E. Barrick, and R.R. Breaker. Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA, 9:644–647, 2003.

[31] W.C. Winkler, A. Nahvi, N. Sudarsan, J.E. Barrick, and R.R. Breaker. An mRNA structure that controls gene

expression by binding Sadenosylmethionine. Nature Structural Biology, 10:701–707, 2003.

[32] J.J. Wyrick, F.C. Holstege, E.G. Jennings, H.C. Causton, D. Shore, M. Grunstein, E.S. Lander, and R.A. Young. Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. Nature, 402:418–421, 1999. T&F Cat # C6847 Chapter: 22 page: 585 date: August 5, 2009

[33] M. Zaki and N. Ramakrishnan. Reasoning about sets using redescription mining. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'05), Chicago, IL, 364–373, 2005.

[34] L. Zhao, M. Zaki, and N. Ramakrishnan. BLOSOM: A framework for mining arbitrary Boolean expressions over attribute sets. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2006), Philadelphia, PA, 827–832, 2006.

23 Chapter 23. Data Mining Tools and Techniques for Identification of Biomarkers for Cancer

Abdel-Qader, I. and F. Abu-Amara. 2008. A computer-aided diagnosis system for breast cancer using independent component analysis and fuzzy classifier. Modeling and Simulation in Engineering, 2008, 1–9.

Asyali, M. H., D. Colak, O. Demirkaya, and M. S. Inan. 2006. Gene expression profile classification: A review. Current Bioinformatics, 1:55–73.

Bellman, R.E. 1961. Adaptive Control Processes. Princeton University Press, Princeton, NJ. T&F Cat # C6847 Chapter: 23 page: 613 date: August 5, 2009

Beaulah, S. A., M. Correll, R. E. J. Munro, and J. G. Sheldon. 2008. Addressing informatics challenges in translational research with workflow technology. Drug Discovery Today, 2008 sept; 13(17–18): 771–7.

Boulesteix, A.-L., C. Strobl, T. Augustin, and M. Daumer. 2008. Evaluating microarray-based classifiers: An overview. Cancer Informatics, 6:77–97.

Brown, M. P. S, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences USA, 97(1):262–267.

Editors. 2000. Looking back on the millennium in medicine. New England Journal of Medicine, 342(1):42–49.

Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth. 1996. From Data Mining to Knowledge Discovery: An overview. In: Advances in Knowledge Discovery, and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurasamy, Menlo Park, California, AAAI Press, 1–30.

Fenton, J. J., S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D’Orsi, E. A. Berns, G. Cutter, R. E. Hendrick, W. E. Barlow, and J. G. Elmore. 2007. Influence of computer-aided detection on performance of screening mammography. New England Journal of Medicine, 356(14):1399–1409.

Hall, B. H., M. Ianosi-Irimie, P. Javidian, W. Chen, S. Ganesan, and D.J. Foran. 2008. Computer-assisted assessment of the Human Epidermal Growth Factor Receptor 2 immunohistochemical assay in imaged histologic sections using a membrane isolation algorithm and quantitative analysis of positive control. BMC Medical Imaging, 8:11. DOI:10.1186/1471-23428-11.

Kerlikowske K., L. Ichikawa, D. L. Miglioretti, D. S. Buist, P. M. Vacek, R. Smith-Bindman, B. Yankaskas, P. A. Carney, and R. Ballard-Barbash. 2007. Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk. Journal of the National Cancer Institute, 99(5):386–395.

Kononen J, Bubendorf L, Kallioniemi A, Ba¨rlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, and Kallioniemi O.P. 1998. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nature Medicine, 4(7):844–847.

Liu J. J., G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling. 2005. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics. 21(11):2691–2697. T&F Cat # C6847 Chapter: 23 page: 614 date: August 5, 2009

Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Radhakrishnan, R. 2008. Tissue microarray—a high-throughput molecular analysis in head and neck cancer. Journal of Oral Pathology and Medicine, 37:166–176.

Ramaswamy S., P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences USA, 98, 15149–15154.

Sheta, W, N. Eltonsy, G. Tourassi, and A. Elmaghraby. 2005. Automated detection of breast cancer from screening mammograms using genetic programming. International Journal of Intelligent Computing and Information Sciences, 5(1):309–318.

Smith R. A. 2007. The evolving role of MRI in the detection and evaluation of breast cancer. New England Journal of Medicine, 356:1362–1364.

Tamayo P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E.

Dmitrovsky, E. S. Lander, and T. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences USA, 96(6):2907–2912.

Tian X, M. R. Aruva, K. Zhang, N. Shanthly, C. A. Cardi, M. L. Thakur and E. Wickstrom. 2007. PET imaging of CCND1 mRNA in human MCF7 estrogen receptor-positive breast cancer xenografts with oncogene-specific [64Cu]chelator-peptide nucleic acid-IGF1 analog radiohybridization probes. Journal of Nuclear Medicine, 48(10):1699–1707.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences USA, 99(10):6567–6572.

Zitova´, B. and J. Flusser. 2003.Image registration methods: A survey. Image and Vision Computing, 21:977–1000.

24 Chapter 24. Cancer Biomarker Prioritization: Assessing the in vivo Impact of in vitro Models by in silico Mining of Microarray Database, Literature, and Gene Annotation

[1] Nass, S.J. and Moses, H.L. (Eds). Cancer Biomarkers: The Promises and Challenges of Improving Detection and Treatment. The National Academies Press, Washington DC, 2007.

[2] Dalton, W.S. and Friend, S.H. Cancer biomarkers—an invitation to the table. Science, 2006, 312(5777), 1165–68.

[3] Hartwell, L., et al. Cancer biomarkers: a systems approach. Nat Biotechnol, 2006, 24(8), 905–8.

[4] Goodsaid, F. and Frueh, F.W. Implementing the U.S. FDA guidance on pharmacogenomic data submissions. Environ Mol Mutagen, 2007, 48(5), 354–58.

[5] Ratner, M. FDA pharmacogenomics guidance sends clear message to industry. Nat Rev Drug Discov, 2005, 4(5), 359.

[6] Biomarkers Definition Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther, 2001, 69(3), 89–95.

[7] Food and Drug Administration. U.S. Department of Health and Human Services, Guidance for industry—Pharmacogenomic data submissions. 2005.

[8] Kelly-Spratt, K.S., et al. A mouse model repository for cancer biomarker discovery. J Proteome Res, 2008, 7(8), 3613–18.

[9] Kuhn, A., Luthi-Carter, R., and Delorenzi, M. Cross-species and cross-platform gene expression studies with the Bioconductorcompliant R package ‘annotationTools’. BMC Bioinform, 2008, 9:26, doi:10.1186/1471-2105-9-26.

[10] Troyanskaya, O.G. Putting microarrays in a context: integrated analysis of diverse biological data. Brief Bioinform, 2005, 6(1), p. 34–43.

[11] Dopazo, J. Functional interpretation of microarray experiments. Omics, 2006, 10(3), 398–410. T&F Cat # C6847 Chapter: 24 page: 625 date: August 5, 2009

[12] Nam, D. and Kim, S.Y. Gene-set approach for expression

pattern analysis. Brief Bioinform, 2008, 189-97.

[13] Al-Shahrour, F., et al. Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. Nucleic Acids Res, 2008, 36(Web Server issue), W341-46.

[14] Rhodes, D.R., et al. Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. Neoplasia, 2007, 9(5), 443-54.

[15] Newman, J.C. and Weiner, A.M. L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol, 2005, 6(9), R81.1-18.

[16] Subramanian, A., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 2005, 102(43), 15545-50.

25 Chapter 25. Biomarker Discovery by Mining Glycomic and Lipidomic Data

[1] Aebersold, R. and M. Mann. Mass spectrometry-based proteomics. Nature, 2003, 422(6928), 198–207.

[2] Varki, A., et al. Essentials of Glycobiology. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1999.

[3] Freeze, H.H. and M. Aebi. Altered glycan structures: the molecular basis of congenital disorders of glycosylation. Curr. Opin. Struct. Biol., 2005, 15(5), 490–498.

[4] Bertozzi, C.R. and L.L. Kiessling. Chemical glycobiology. Science, 2001, 291, 2357–2364.

[5] Buhman, K.K., H.C. Chen, and R.V. Farese, Jr. The enzymes of neutral lipid synthesis. J. Biol. Chem., 2001, 276(44), 40369–40372.

[6] Dell, A. and H.R. Morris. Glycoprotein structure determination by mass spectrometry. Science, 2001, 291(5512), 2351–2356.

[7] Zaia, J. Mass spectrometry of oligosaccharides. Mass Spectrom. Rev., 2004, 23(3), 161–227.

[8] Novotny, M.V. and Y. Mechref. New hyphenated methodologies in highsensitivity glycoprotein analysis. J. Sep. Sci., 2005, 28(15), 1956–1968.

[9] Mechref, Y. and M.V. Novotny. Miniaturized separation techniques in glycomic investigations. J. Chromatogr. B, 2006, 841(1–2), 65–78.

[10] Roberts, L.D., et al. A matter of fat: an introduction to lipidomic profiling methods. J. Chromatogr. B, 2008, 174–181.

[11] Isaac, G., et al. New mass-spectrometry-based strategies for lipids. Genet. Eng. (NY), 2007, 28, 129–157.

[12] Stevens, J., et al. Glycan microarray technologies: tools to survey host specificity of influenza viruses. Nat. Rev. Microbiol., 2006, 4, 857–864. T&F Cat # C6847 Chapter: 25 page: 644 date: August 5, 2009

[13] Dyukova, V.I., et al. Hydrogel glycan microarrays.

Anal. Biochem., 2005, 347(1), 94–105.

[14] Feizi, T., et al. Carbohydrate microarrays—a new set of technologies at the frontiers of glycomics. Curr. Opin. Struct. Biol., 2003, 13, 637–645.

[15] Xia, B., et al. Versatile fluorescent derivatization of glycans for glycomic analysis. Nat. Methods, 2005, 2(11), 845–850.

[16] Feng, L. Probing lipid-protein interactions using lipid microarrays. Prost. & Other Lipid Mediators, 2005, 77(1–4), 158–167.

[17] Kanter, J.L., et al. Lipid microarrays identify key mediators of autoimmune brain inflammation. Nat. Med., 2006, 12(1), 138–143.

[18] Miyamoto, S. Clinical applications of glycomic approaches for the detection of cancer and other diseases. Curr. Opin. Mol. Ther., 2006, 8(6), 507–513.

[19] Wiest, M.M. and S.M. Watkins. Biomarker discovery using highdimensional lipid analysis. Curr. Opin. Lipidol., 2007, 18(2), 181–186.

[20] Gross, R.W. and X. Han. Unlocking the complexity of lipids: using lipidomics to identify disease mechanisms, biomarkers and treatment efficacy. Future Lipidol., 2006, 1(5), 539–547.

[21] Wuhrer, M. Glycosylation profiling in clinical proteomics: heading for glycan biomarkers. Expert Rev. Proteomics, 2007, 4(2), 135–136.

[22] von der Lieth, C.-W., T. Lu¨tteke, and M. Frank. The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. Biochim. Biophys. Acta (BBA), St. Louis, MO, USA. 2006, 1760(4), 568–577.

[23] Perez, S. and B. Mulloy. Prospects for glycoinformatics. Curr. Opin. Struct. Biol., 2005, 15(5), 517–524.

[24] Fahy, E., et al. Bioinformatics for lipidomics. In Methods in Enzymology. Academic Press, 2007, 247–273.

[25] Dwek, R.A. Glycobiology: toward understanding the function of sugars. Chem. Rev., 1996, 96(2), 683–720.

[26] Mechref, Y. and M.V. Novotny. Structural investigations of glycoconjugates at high sensitivity. Chem. Rev., 2002, 102(2), 321–369.

[27] Nishimura, S.-I., et al. High-throughput protein glycomics: combined use of chemoselective glycoblotting and MALDI-TOF/TOF mass spectrometry. Angew Chem. Int. Ed. Engl., 2004, 44, 91–96. T&F Cat # C6847 Chapter: 25 page: 645 date: August 5, 2009 Data

[28] Ressom, H.W., et al. Analysis of MALDI-TOF mass spectrometry data for discovery of peptide and glycan biomarkers of hepatocellular carcinoma. J. Proteome Res., 2008, 7(2), 603–610.

[29] Kyselova, Z., et al. Breast cancer diagnosis and prognosis through quantitative measurements of serum glycan profiles. Clin. Chem., 2008, 54(7), 1166–1175.

[30] Kyselova, Z., et al. Alterations in the serum glycome due to metastatic prostate cancer. J. Proteome Res., 2007, 6(5), 1822–1832.

[31] Jang-Lee, J., et al. Glycomic profiling of cells and tissues by mass spectrometry: fingerprinting and sequencing methodologies. Methods Enzymol., 2006, 415, 59–86.

[32] Ethier, M., D. Figeys, and H. Perreault. N-glycosylation analysis using the StrOligo algorithm. Methods Mol. Biol., 2006, 328, 187–197.

[33] Mechref, Y., N.V. Novotny, and C. Krishnan. Structural characterization of oligosaccharides using MALDI-TOF/TOF tandem mass spectrometry. Anal. Chem., 2003, 75(18), 4895–4903.

[34] Goldberg, D., et al. Automatic determination of O-glycan structure from fragmentation pectra. J. Proteome Res., 2006, 5(6), 1429–1434.

[35] Goldberg, D., et al. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. Proteomics, 2005, 5(4), 865–875.

[36] Ceroni, A., et al. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. J. Proteome Res., 2008, 7(4), 1650– 1659.

[37] Ethier, M., et al. Global and site-specific detection

of human integrin α5β1 glycosylation using tandem mass spectrometry and the StrOligo algorithm. Rapid Commun. Mass Spectrom., 2005, 19(5), 721-727.

[38] Lohmann, K.K. and C.-W. von der Lieth. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. Nucl. Acids Res., 2004, 32(suppl 2), W261- W266.

[39] Ethier, M., et al. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. Rapid Commun. Mass Spectrom., 2003, 17(24), 2713-2720.

[40] Irungu, J., et al. Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID. Anal. Chem., 2007, 79(8), 3065-3074. T&F Cat # C6847 Chapter: 25 page: 646 date: August 5, 2009

[41] Lapadula, A.J., et al. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. Anal. Chem., 2005, 77(19), 6271-6279.

[42] Mechref, Y., M.V. Novotny, and C. Krishnan. Structural characterization of oligosaccharides using Maldi-TOF/TOF tandem mass spectrometry. Anal. Chem., 2003, 75(18), 4895-4903.

[43] Tang, H., Y. Mechref, and M.V. Novotny. Automated interpretation of MS/MS spectra of oligosaccharides. Bioinformatics, 2005, 21(suppl 1), i431-i439.

[44] Ashline, D.J., et al. Carbohydrate structural isomers analyzed by sequential mass spectrometry. Anal. Chem., 2007, 79(10), 3830-3842.

[45] Takegawa, Y., et al. Separation of isomeric 2-aminopyridine derivatized N-glycans and N-glycopeptides of human serum immunoglobulin G by using a zwitterionic type of hydrophilic-interaction chromatography. J. Chromatogr. A, 2006, 1113(1-2), 177-181.

[46] Zhuang, Z., et al. Electrophoretic analysis of N-glycans on microfluidic devices. Anal. Chem., 2007, 79(18), 7170-7175.

[47] Isailovic, D., et al. Profiling of human serum glycans associated with liver cancer and cirrhosis by IMS/MS. J.

Proteome Res., 2008, 7(3), 1109–1117.

[48] Devakumar, A., et al. Laser-induced photofragmentation of neutral and acidic glycans inside an ion-trap mass spectrometer. Rapid Commun. Mass Spectrom., 2007, 21(8), 1452–1460.

[49] Yin, W., et al. A computational approach for the identification of sitespecific protein glycosylations through ion-trap mass spectrometry. In Proceedings of RECOMB Satellite Conferences on: Systems Biology and Computational Proteomics. The third RECOMB satellite meeting on Proteomics, Lecture Notes in Bioinformatics, Springer, 2007, 4532(96–107).

[50] Goldberg, D., et al. Automated N-glycopeptide identification using a combination of singleand tandem-MS. J. Proteome Res., 2007, 6(10), 3995–4005.

[51] Mamitsuka, H. Informatic innovations in glycobiology: relevance to drug discovery. Drug Discov. Today, 2008, 13, 118–23.

[52] Aoki-Kinoshita, K.F. An introduction to bioinformatics for glycomics research. PLoS Computation. Biol., 2008, 4(5), e1000075.

[53] Hizukuri, Y., et al. Extraction of leukemia specific glycan motifs in humans by computational glycomics. Carbohydr. Res., 2005, 340(14), 2270–2278. T&F Cat # C6847 Chapter: 25 page: 647 date: August 5, 2009 Data

[54] Yamanishi, Y., F. Bach, and J.-P. Vert. Glycan classification with tree kernels. Bioinformatics, 2007, 23(10), 1211–1216.

[55] Aoki, K.F., et al. Application of a new probabilistic model for recognizing complex patterns in glycans. Bioinformatics, 2004, 20(suppl 1), i6–i14.

[56] Aoki-Kinoshita, K.F., et al. ProfilePSTMM: capturing tree-structure motifs in carbohydrate sugar chains. Bioinformatics, 2006, 22(14), e25–e34.

[57] Fahy, E., et al. A comprehensive classification system for lipids. J. Lipid Res., 2005, 46(5), 839–862.

[58] Han, X. and R.W. Gross. Shotgun lipidomics: multidimensional MS analysis of cellular lipidomes. Expert Rev. Proteomics, 2005, 2(2), 253–264.

[59] Milne, S., et al. Lipidomics: an analysis of cellular lipids by ESI-MS. Methods, 2006, 39(2), 92–103.

[60] Brugger, B., et al. Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem mass spectrometry. Proc. Natl. Acad. Sci. USA, 1997, 94(6), 2339–2344.

[61] Knochenmuss, R. Ion formation mechanisms in UV-MALDI. The Analysts, 2006, 131, 966–986.

[62] Cornett, D.S., et al. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. Nat. Methods, 2007, 4(10), 828–833.

[63] Walch, A., et al. MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. Histochem. Cell Biol., 2008, 130(3), 421–434.

[64] Wenk, M.R., et al. Phosphoinositide profiling in complex lipid mixtures using electrospray ionization mass spectrometry. Nat. Biotech., 2003, 21(7), 813–817.

[65] Petkovic, M., et al. Detection of individual phospholipids in lipid mixtures by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry: Phosphatidylcholine prevents the detection of further species. Anal. Biochem., 2001, 289(2), 202–216.

[66] Watanabe, K., E. Yasugi, and M. Oshima. How to search the glycolipid data in LIPIDBANK for Web: the newly developed lipid database. Japan Trend Glycosci. Glycotechnol., 2000, 12, 175–184.

[67] Houjou, T., et al. A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry. Rapid Commun. Mass Spectrom., 2005, 19(5), 654–666. T&F Cat # C6847 Chapter: 25 page: 648 date: August 5, 2009

[68] Song, H., et al. Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. J. Am. Soc. Mass Spectrom., 2007, 18(10), 1848–1858.

[69] Haimi, P., et al. Software tools for analysis of mass spectrometric lipidome data. Anal. Chem., 2006, 78(24),

8324-8331.

[70] Yetukuri, L., et al. Informatics and computational strategies for the study of lipids. Mol. BioSyst., 2008, 4, 121-127.

[71] Yetukuri, L., et al. Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. BMC Syst Biol., 2007, 1(12), 12-26.

[72] Tang, H., et al. A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics, 2006, 22(14), e481-e488.

[73] Lu, P., et al. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotech., 2007, 25(1), 117-124.

[74] Ressom, H.W., et al. Analysis of MALDI-TOF mass spectrometry data for detection of glycan biomarkers. Pac. Symp. Biocomput., 2008, 216-227.

[75] Welsh, J.B., et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA, 2001, 98(3), 1176-1181.

[76] Ramaswamy, S., et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA, 2001, 98(26), 15149-15154.

[77] Zhang, Z., et al. Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res., 2004, 64(16), 5882-5890.

[78] Chuang, H.-Y., et al. Network-based classification of breast cancer metastasis. Mol. Syst. Biol., 2007, 3, 140-149.

26 Chapter 26. Data Mining Chemical Structures and Biological Data

Barnard, J.M. and Downs, G.M. 1992. Clustering of chemical structures on the basis of two-dimensional similarity measures. J. Chem. Inf. Comput. Sci., 32:644–49.

Bredel, M. and Jacoby, E. 2004. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat. Rev. Genet., 5:262–75.

Blower, P.E., Cross, K.P., Fligner, M.A., Myatt, G.J., Verducci, J.S., and Yang, C. 2004. Systematic analysis of large screening sets in drug discovery. Curr. Drug Discov. Technol., 42:393–404.

Blower, P.E., Yang, C., Fligner, M.A., Verducci, J.S., Yu, L., Richman, S., and Weinstein, J.N. 2002. Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. Pharmacogenomics J., 2:259–71.

Gasteiger, J. and Engel T. 2003. Chemoinformatics: A Textbook. Wiley-VCH, Weinheim.

Giaever, G., Flaherty, P., Kumm, J., Proctor, M., Nislow, C., Jaramillo, D.F., Chu, A.M., Jordan, M.I., Arkin, A.P., and Davis, R.W. 2004. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. Proc. Natl. Acad. Sci. USA, 101:793–98.

Hawkins, D.M., Young, S., and Rusinko, A. 1997. Analysis of large structureactivity data set using recursive partitioning. Quant. Strct.-Act. Relat., 16:296–302.

Huang, Y., Blower, P.E., Yang, C., Barbacioru, C., Dai, Z., Zhang, Y., Xiao, J.J., Chan, K.K., and Sade´e, W. 2005. Correlating gene expression with chemical scaffolds of cytotoxic agents: ellipticines as substrates and inhibitors of MDR1. Pharmacogenomics J., 5:112–25.

Leach, A.R. and Gillet, V.J. 2007. An Introduction to Chemoinformatics. Springer, Dordrecht, The Netherlands.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J. et al. 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313:1929–35.

Liu, K., Feng, J., and Young, S.S. 2005. PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. J. Chem. Inf. Model., 45:515–22. T&F Cat # C6847 Chapter: 26 page: 687 date: August 5, 2009

Potti, A., Dressman, H.K., Bild, A., Riedel, R.F., Chan, G., Sayer, R., Cragun, J. et al. 2006. Genomic signatures to guide the use of chemotherapeutics. Nat. Med., 12:1294–300.

Wallqvist, A., Rabow, A.A., Shoemaker, R.H., Sausville, E.A., and Covell, D.G. 2003. Linking the growth inhibition response from the National Cancer Institute's anticancer screen to gene expression levels and other molecular target data. Bioinformatics. 19:2212–24

Weinstein, J.N. 2006. Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. Mol. Cancer Ther., 5:2601–5.

Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J. Jr., Kohn, K.W., Fojo, T. et al. 1997. An information-intensive approach to the molecular pharmacology of cancer. Science, 275:343–49.