FUNDAMENTALS OF DATA MINING IN GENOMICS AND PROTEOMICS

FUNDAMENTALS OF DATA MINING IN GENOMICS AND PROTEOMICS

Edited by
Werner Dubitzky
University of Ulster, Coleraine, Northern Ireland

Martin Granzow
Quantiom Bioinformatics GmbH & Co. KG, Weingarten/Baden, Germany

Daniel Berrar
University of Ulster, Coleraine, Northern Ireland

Library of Congress Control Number: 2006934109

ISBN-13: 978-0-387-47508-0 e-ISBN-13: 978-0-387-47509-7
ISBN-10: 0-387-47508-7 e-ISBN-10: 0-387-47509-5

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Preface

As natural phenomena are being probed and mapped in ever-greater detail, scientists in genomics and proteomics are facing an exponentially growing volume of increasingly complex-structured data, information, and knowledge. Examples include data from microarray gene expression experiments, bead-based and microfluidic technologies, and advanced high-throughput mass spectrometry. A fundamental challenge for life scientists is to explore, analyze, and interpret this information effectively and efficiently. To address this challenge, traditional statistical methods are being complemented by methods from data mining, machine learning and artificial intelligence, visualization techniques, and emerging technologies such as Web services and grid computing.

There exists a broad consensus that sophisticated methods and tools from statistics and data mining are required to address the growing data analysis and interpretation needs in the life sciences. However, there is also a great deal of confusion about the arsenal of available techniques and how these should be used to solve concrete analysis problems. Partly this confusion is due to a lack of mutual understanding caused by the different concepts, languages, methodologies, and practices prevailing within the different disciplines.

A typical scenario from pharmaceutical research should illustrate some of the issues. A molecular biologist conducts nearly one hundred experiments examining the toxic effect of certain compounds on cultured cells using a microarray gene expression platform. The experiments include different compounds and doses and involves nearly 20 000 genes. After the experiments are completed, the biologist presents the data to the bioinformatics department and briefly explains what kind of questions the data is supposed to answer. Two days later the biologist receives the results which describe the output of a cluster analysis separating the genes into groups of activity and dose. While the groups seem to show interesting relationships, they do not directly address the questions the biologist has in mind. Also, the data sheet accompanying the results shows the original data but in a different order and somehow transformed. Discussing this with the bioinformatician again it turns out that what

the biologist wanted was not clustering (*automatic* classification or *automatic* class prediction) but *supervised* classification or *supervised* class prediction.

One main reason for this confusion and lack of mutual understanding is the absence of a conceptual platform that is common to and shared by the two broad disciplines, life science and data analysis. Another reason is that data mining in the life sciences is different to that in other typical data mining applications (such as finance, retail, and marketing) because many requirements are fundamentally different. Some of the more prominent differences are highlighted below.

A common theme in many genomic and proteomic investigations is the need for a detailed understanding (descriptive, predictive, explanatory) of genome- and proteome-related entities, processes, systems, and mechanisms. A vast body of knowledge describing these entities has been accumulated on a staggering range of life phenomena. Most conventional data mining applications do not have the requirement of such a deep understanding and there is nothing that compares to the global knowledge base in the life sciences.

A great deal of the data generated in genomics and proteomics is generated in order to analyze and interpret them in the context of the questions and hypotheses to be answered and tested. In many classical data mining scenarios, the data to be analyzed are generated as a "by-product" of an underlying business process (e.g., customer relationship management, financial transactions, process control, Web access log, etc.). Hence, in the conventional scenario there is no notion of question or hypothesis at the point of data generation.

Depending on what phenomenon is being studied and the methodology and technology used to generate data, genomic and proteomic data structures and volumes vary considerably. They include temporally and spatially resolved data (e.g., from various imaging instruments), data from spectral analysis, encodings for the sequential and spatial representation of biological macromolecules and smaller chemical and biochemical compounds, graph structures, and natural language text, etc. In comparison, data structures encountered in typical data mining applications are simple.

Because of ethical constraints and the costs and time involved to run experiments, most studies in genomics and proteomics create a modest number of observation points ranging from several dozen to several hundreds. The number of observation points in classical data mining applications ranges from thousands to millions. On the other hand, modern high-throughput experiments measure several thousand variables per observation, much more than encountered in conventional data mining scenarios.

By definition, research and development in genomics and proteomics is subject to constant change – new questions are being asked, new phenomena are being probed, and new instruments are being developed. This leads to frequently changing data processing pipelines and workflows. Business processes in classical data mining areas are much more stable. Because solutions will be in use for a long time, the development of complex, comprehensive, and

expensive data mining applications (such as data warehouses) is readily justified.

Genomics and proteomics are intrinsically "global" – in the sense that hundreds if not thousands of databases, knowledge bases, computer programs, and document libraries are available via the Internet and are used by researchers and developers throughout the world as part of their day-to-day work. The information accessible through these sources form an intrinsic part of the data analysis and interpretation process. No comparable infrastructure exists in conventional data mining scenarios.

This volume presents state of the art analytical methods to address key analysis tasks that data from genomics and proteomics involve. Most importantly, the book will put particular emphasis on the common caveats and pitfalls of the methods by addressing the following questions: What are the requirements for a particular method? How are the methods deployed and used? When should a method not be used? What can go wrong? How can the results be interpreted? The main objectives of the book include:

- To be acceptable and accessible to researchers and developers both in life science and computer science disciplines – it is therefore necessary to express the methodology in a language that practitioners in both disciplines understand;
- To incorporate fundamental concepts from both conventional statistics as well as the more exploratory, algorithmic and computational methods provided by data mining;
- To take into account the fact that data analysis in genomics and proteomics is carried out against the backdrop of a huge body of existing formal knowledge about life phenomena and biological systems;
- To consider recent developments in genomics and proteomics such as the need to view biological entities and processes as systems rather than collections of isolated parts;
- To address the current trend in genomics and proteomics towards increasing computerization, for example, computer-based modeling and simulation of biological systems and the data analysis issues arising from large-scale simulations;
- To demonstrate where and how the respective methods have been successfully employed and to provide guidelines on how to deploy and use them;
- To discuss the advantages and disadvantages of the presented methods, thus allowing the user to make an informed decision in identifying and choosing the appropriate method and tool;
- To demonstrate potential caveats and pitfalls of the methods so as to prevent any inappropriate use;
- To provide a section describing the formal aspects of the discussed methodologies and methods;

- To provide an exhaustive list of references the reader can follow up to obtain detailed information on the approaches presented in the book;
- To provide a list of freely and commercially available software tools.

It is hoped that this volume will (*i*) foster the understanding and use of powerful statistical and data mining methods and tools in life science as well as computer science and (*ii*) promote the standardization of data analysis and interpretation in genomics and proteomics.

The approach taken in this book is conceptual and practical in nature. This means that the presented data-analytical methodologies and methods are described in a largely non-mathematical way, emphasizing an information-processing perspective (input, output, parameters, processing, interpretation) and conceptual descriptions in terms of mechanisms, components, and properties. In doing so, the reader is not required to possess detailed knowledge of advanced theory and mathematics. Importantly, the merits and limitations of the presented methodologies and methods are discussed in the context of "real-world" data from genomics and proteomics. Alternative techniques are mentioned where appropriate. Detailed guidelines are provided to help practitioners avoid common caveats and pitfalls, e.g., with respect to specific parameter settings, sampling strategies for classification tasks, and interpretation of results. For completeness reasons, a short section outlining mathematical details accompanies a chapter if appropriate. Each chapter provides a rich reference list to more exhaustive technical and mathematical literature about the respective methods.

Our goal in developing this book is to address complex issues arising from data analysis and interpretation tasks in genomics and proteomics by providing what is simultaneously a *design blueprint*, *user guide*, and *research agenda* for current and future developments in the field.

As design blueprint, the book is intended for the practicing professional (researcher, developer) tasked with the analysis and interpretation of data generated by high-throughput technologies in genomics and proteomics, e.g., in pharmaceutical and biotech companies, and academic institutes.

As a user guide, the book seeks to address the requirements of scientists and researchers to gain a basic understanding of existing concepts and methods for analyzing and interpreting high-throughput genomics and proteomics data. To assist such users, the key concepts and assumptions of the various techniques, their conceptual and computational merits and limitations are explained, and guidelines for choosing the methods and tools most appropriate to the analytical tasks are given. Instead of presenting a complete and intricate mathematical treatment of the presented analysis methodologies, our aim is to provide the users with a clear understanding and practical know-how of the relevant concepts and methods so that they are able to make informed and effective choices for data preparation, parameter setting, output post-processing, and result interpretation and validation.

As a research agenda, this volume is intended for students, teachers, researchers, and research managers who want to understand the state of the art of the presented methods and the areas in which gaps in our knowledge demand further research and development. To this end, our aim is to maintain the readability and accessibility throughout the chapters, rather than compiling a mere reference manual. Therefore, considerable effort is made to ensure that the presented material is supplemented by rich literature cross-references to more foundational work.

In a quarter-length course, one lecture can be devoted to two chapters, and a project may be assigned based on one of the topics or techniques discussed in a chapter. In a semester-length course, some topics can be covered in greater depth, covering – perhaps with the aid of an in-depth statistics/data mining text – more of the formal background of the discussed methodology. Throughout the book concrete suggestions for further reading are provided.

Clearly, we cannot expect to do justice to all three goals in a single book. However, we do believe that this book has the potential to go a long way in bridging a considerable gap that currently exists between scientists in the field of genomics and proteomics on one the hand and computer scientists on the other hand. Thus, we hope, this volume will contribute to increased communication and collaboration across the disciplines and will help facilitate a consistent approach to analysis and interpretation problems in genomics and proteomics in the future.

This volume comprises 12 chapters, which follow a similar structure in terms of the main sections. The centerpiece of each chapter represents a case study that demonstrates the use – and misuse – of the presented method or approach. The first chapter provides a general introduction to the field of data mining in genomics and proteomics. The remaining chapters are intended to shed more light on specific methods or approaches.

The second chapter focuses on study design principles and discusses replication, blocking, and randomization. While these principles are presented in the context of microarray experiments, they are applicable to many types of experiments.

Chapter 3 addresses data pre-processing in cDNA and oligonucleotide microarrays. The methods discussed include background intensity correction, data normalization and transformation, how to make gene expression levels comparable across different arrays, and others.

Chapter 4 is also concerned with pre-processing. However, the focus is placed on high-throughput mass spectrometry data. Key topics include baseline correction, intensity normalization, signal denoising (e.g., via wavelets), peak extraction, and spectra alignment.

Data visualization plays an important role in exploratory data analysis. Generally, it is a good idea to look at the distribution of the data prior to analysis. Chapter 5 revolves around visualization techniques for high-dimensional data sets, and puts emphasis on multi-dimensional scaling. This technique is illustrated on mass spectrometry data.

Chapter 6 presents the state of the art of clustering techniques for discovering groups in high-dimensional data. The methods covered include hierarchical and k-means clustering, self-organizing maps, self-organizing tree algorithms, model-based clustering, and cluster validation strategies, such as functional interpretation of clustering results in the context of microarray data.

Chapter 7 addresses the important topics of feature selection, feature weighting, and dimension reduction for high-dimensional data sets in genomics and proteomics. This chapter also includes statistical tests (parametric or nonparametric) for assessing the significance of selected features, for example, based on random permutation testing.

Since data sets in genomics and proteomics are usually relatively small with respect to the number of samples, predictive models are frequently tested based on resampled data subsets. Chapter 8 reviews some common data resampling strategies, including n-fold cross-validation, leave-one-out cross-validation, and repeated hold-out method.

Chapter 9 discusses support vector machines for classification tasks, and illustrates their use in the context of mass spectrometry data.

Chapter 10 presents graphs and networks in genomics and proteomics, such as biological networks, pathways, topologies, interaction patterns, gene-gene interactome, and others.

Chapter 11 concentrates on time series analysis in genomics. A methodology for identifying important predictors of time-varying outcomes is presented. The methodology is illustrated in a study aimed at finding mutations of the human immunodeficiency virus that are important predictors of how well a patient responds to a drug regimen containing two different antiretroviral drugs.

Automated extraction of information from biological literature promises to play an increasingly important role in text-based knowledge discovery processes. This is particularly important for high-throughput approaches such as microarrays and high-throughput proteomics. Chapter 12 addresses knowledge extraction via text mining and natural language processing.

Finally, we would like to acknowledge the excellent contributions of the authors and Alice McQuillan for her help in proofreading.

Coleraine, Northern Ireland, and Weingarten, Germany

Werner Dubitzky
Martin Granzow
Daniel Berrar

The following list shows the symbols or abbreviations for the most commonly occurring quantities/terms in the book. In general, uppercase boldfaced letters such as $\mathbf{X}$ refer to matrices. Vectors are denoted by lowercase boldfaced letters, e.g., $\mathbf{x}$, while scalars are denoted by lowercase italic letters, e.g., x.

List of Abbreviations and Symbols

ACE	Average (test) classification error
ANOVA	Analysis of variance
ARD	Automatic relevance determination
AUC	Area under the curve (in ROC analysis)
BACC	Balanced accuracy (average of sensitivity and specificity)
BACC	Balanced accuracy
bp	Base pair
CART	Classification and regression tree
CV	Cross-validation
Da	Daltons
DDWT	Decimated discrete wavelet transform
ESI	Electrospray ionization
EST	Expressed sequence tag
ETA	Experimental treatment assignment
FDR	False discovery rate
FLD	Fisher's linear discriminant
FN	False negative
FP	False positive
FPR	False positive rate
FWER	Family-wise error rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
ICA	Independent component analysis
IE	Information extraction
IQR	Interquartile range
IR	Information retrieval
LOOCV	Leave-one-out cross-validation
MALDI	Matrix-assisted laser desorption/ionization
MDS	Multidimensional scaling
MeSH	Medical Subject Headings
MM	Mismatch
MS	Mass spectrometry
m/z	Mass-over-charge
NLP	Natural language processing
NPV	Negative predictive value
PCA	Principal component analysis
PCR	polymerase chain reaction

PCR	Polymerase chain reaction
PLS	Partial least squares
PM	Perfect match
PPV	Positive predictive value
RLE	Relative log expression
RLR	Regularized logistic regression
RMA	Robust multi-chip analysis
S2N	Signal-to-noise
SAGE	Serial analysis of gene expression
SAM	Significance analysis of gene expression
SELDI	Surface-enhance laser desorption/ionization
SOM	Self-organizing map
SOTA	Self-organizing tree algorithm
SSH	Suppression substractive hybridization
SVD	Singular value decomposition
SVM	Support vector machine
TIC	Total ion current
TN	True negative
TOF	Time-of-flight
TP	True positive
UDWT	Undecimated discrete wavelet transform
VSN	Variance stabilization normalization
$\#(\cdot)$	Counts; the number of instances satisfying the condition in $(\cdot)$
$\bar{\mathbf{x}}$	The mean of all elements in $\mathbf{x}$
χ^2	Chi-square statistic
ϵ	Observed error rate
$\epsilon_{.632}$	Estimate for the classification error in the .632 bootstrap
$\hat{y}_i$	Predicted value for y_i (i.e., predicted class label for case $\mathbf{x}_i$)
$\neg y$	Not y
Σ	Covariance
τ	True error rate
$\mathbf{x}'$	Transpose of vector $\mathbf{x}$
D	Data set
$d(x, y)$	Distance between x and y
$E(X)$	Expectation of a random variable X
$\langle k \rangle$	Average of k
L_i	i^{th} learning set
$\Re$	Set of real numbers
T_i	i^{th} test set
TR_{ij}	Training set of the i^{th} external and j^{th} internal loop
V_{ij}	Validation set of the i^{th} external and j^{th} internal loop
v_i	i^{th} vertex in a network

Contents

List of Contributors

Keith A. Baggerly
Department of Biostatistics and Applied Mathematics, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA.
kabagg@wotan.mdacc.tmc.edu

Oliver Bembom
Division of Biostatistics, University of California, Berkeley, CA 94720-7360, USA.
bembom@berkeley.edu

Daniel Berrar
Systems Biology Research Group, University of Ulster, Northern Ireland, UK.
dp.berrar@ulster.ac.uk

Benjamin M. Bolstad
Department of Statistics, University of California, Berkeley, CA 94720-3860, USA.
bmb@bmbolstad.com

Kevin R. Coombes
Department of Biostatistics and Applied Mathematics, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA.
krc@odin.mdacc.tmc.edu

Joaquín Dopazo
Department of Bioinformatics, Centro de Investigación Príncipe Felipe, E46013, Valencia, Spain.
jdopazo@cipf.es

Werner Dubitzky
Systems Biology Research Group, University of Ulster, Northern Ireland, UK.
w.dubitzky@ulster.ac.uk

Martin Granzow
quantiom bioinformatics GmbH & Co. KG, Ringstrasse 61, D-76356 Weingarten, Germany.
martin.granzow@quantiom.de

Jaroslaw Harezlak
Harvard School of Public Health, Boston, MA 02115, USA.
jharezla@hsph.harvard.edu

Milos Hauskrecht
Department of Computer Science, and Intelligent Systems Program, and Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA.
milos@cs.pitt.edu

Robert Hoffmann
Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA.
hoffmann@cbio.mskcc.org

Peter Johansson
Computational Biology and Biological Physics Group, Department of Theoretical Physics, Lund University, SE-223 62, Lund, Sweden.
peter@thep.lu.se

Kathleen F. Kerr
Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.
katiek@u.washington.edu

Xiaochun Li
Dana Farber Cancer Institute, Boston, Massachusetts, USA, and Harvard School of Public Health, Boston, MA 02115, USA.
xiaochun@jimmy.harvard.edu

James Lyons-Weiler
Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA.
lyonsweilerj@upmc.edu

Jeffrey S. Morris
Department of Biostatistics and Applied Mathematics, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA.
jeffmo@wotan.mdacc.tmc.edu

Richard Pelikan
Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA.
pelikan@cs.pitt.edu

Maya L. Petersen
Division of Biostatistics, University of California, Berkeley, CA 94720-7360, USA.
mayaliv@berkeley.edu

Markus Ringnér
Computational Biology and Biological Physics Group, Department of Theoretical Physics, Lund University, SE-223 62, Lund, Sweden.
markus@thep.lu.se

Carlos Rodríguez-Caso
ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain.
carlos.rodriguez@upf.edu

Richard Simon
National Cancer Institute, Rockville, MD 20852, USA.
rsimon@mail.nih.gov

Ricard V. Solé
ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain, and Santa Fe Institute, 1399 Hyde Park Road, NM 87501, USA.
ricard.sole@upf.edu

Michal Valko
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260, USA.
michal@cs.pitt.edu

Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley, CA 94720-7360, USA.
laan@stat.berkeley.edu