A PRACTICAL APPROACH TO MICROARRAY DATA ANALYSIS

A PRACTICAL APPROACH TO MICROARRAY DATA ANALYSIS

edited by

Daniel P. Berrar
School of Biomedical Sciences
University of Ulster at Coleraine, Northern Ireland

Werner Dubitzky
Faculty of Life and Health Science
and Faculty of Informatics
University of Ulster at Coleraine, Northern Ireland

Martin Granzow
4T2consulting
Weingarten, Germany

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47815-3
Print ISBN: 1-4020-7260-0

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

Contents

Acknowledgements

The editors would like to thank the contributing authors for their excellent work. Furthermore, the editors would like to thank Joanne Tracy and Dianne Wuori from Kluwer Academic Publishers for their help and support in editing this volume.

Preface

In the past several years, DNA microarray technology has attracted tremendous interest in both the scientific community and in industry. With its ability to simultaneously measure the activity and interactions of thousands of genes, this modern technology promises unprecedented new insights into mechanisms of living systems. Currently, the primary applications of microarrays include gene discovery, disease diagnosis and prognosis, drug discovery (pharmacogenomics), and toxicological research (toxicogenomics).

Typical *scientific tasks* addressed by microarray experiments include the identification of coexpressed genes, discovery of sample or gene groups with similar expression patterns, identification of genes whose expression patterns are highly differentiating with respect to a set of discerned biological entities (e.g., tumor types), and the study of gene activity patterns under various stress conditions (e.g., chemical treatment). More recently, the discovery, modeling, and simulation of regulatory gene networks, and the mapping of expression data to metabolic pathways and chromosome locations have been added to the list of scientific tasks that are being tackled by microarray technology.

Each scientific task corresponds to one or more so-called *data analysis tasks.* Different types of scientific questions require different sets of data analytical techniques. Broadly speaking, there are two classes of elementary data analysis tasks, *predictive modeling* and *pattern-detection.* Predictive modeling tasks are concerned with learning a classification or estimation function, whereas pattern-detection methods screen the available data for interesting, previously unknown regularities or relationships.

A plethora of sophisticated methods and tools have been developed to address these tasks. However, each of these methods is characterized by a set of idiosyncratic requirements in terms of data pre-processing, parameter configuration, and result evaluation and interpretation. To optimally design and analyze microarray experiments, researchers and developers need a

sufficient overview of existing methodologies and tools and a basic understanding of how to apply them.

We believe that one significant barrier to the widespread effective and efficient use of microarray analysis methods and tools is a lack of a clear understanding of how such techniques are used, what their merits and limitations are, and what obstacles are involved in deploying them. Our goal in developing this book was to address this issue, 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 life scientists, statisticians, computer experts, technology developers, managers, and other professionals who will be tasked with developing, deploying, and using microarray technology including the necessary computational infrastructure and analytical tools.

As a user guide, the book seeks to address the requirement of scientists and researchers to gain a basic understanding of microarray analysis methodologies and tools. For these users, we seek to explain the key concepts and assumptions of the various techniques, their conceptual and computational merits and limitations, and give guidelines for choosing the methods and tools most appropriate for the analytical task at hand. Our emphasis is not on a complete and intricate mathematical treatment of the presented analysis methodologies. Instead, we aim at providing the users with a clear understanding and practical know-how of the relevant 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. For methodologies where free software exists we will also provide practical tips for obtaining and using the tools.

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 methodologies and the areas in which gaps in our knowledge demand further research and development. To this end, our aim was to maintain the readability and accessibility of a textbook throughout the chapters, rather than compiling a mere reference manual. Therefore, considerable effort was made to ensure that the presented material, which stresses the applied aspects of microarray analysis, is supplemented by rich literature cross-references to more foundational work.

Clearly, we cannot expect to do justice to all three goals in a single book. However, we do believe that we have succeeded in taking useful steps toward each goal. In doing so, we hope to advance the understanding of both the methodologies and tools needed to analyze microarray data, and the implications for future developments of microarray technology and its support technologies.

The design and subsequent analytical examination of microarray experiments rests on the scientific expertise of the experimenters, their knowledge of the relevant microarray technology and experimental protocols, and their understanding of analysis methods and tools. The available machinery of microarray analysis methods ranges from classical statistical approaches, to machine learning techniques and to methods from artificial intelligence. Hence, the preparation of this book must draw upon the experts from many diverse subfields in mathematics and computer science. In developing this volume, we have assembled a distinguished set of authors, each recognized as an authority in one or more of these fields. We have asked these authors to present a selected set of state-of-the-art methodologies and tools for analyzing microarrays from a highly practical, user-oriented perspective, emphasizing the how-to aspects of the presented techniques. To support the research agenda of this book, we have also asked the authors to identify where future developments are likely to take place and to provide a rich set of pointers to theoretical works underpinning the presented methods. The result, we hope, is a book that will be valuable for a long time, as summary of where we are, as a practical user guide for making informed choices on actual microarray analysis projects, and as roadmap for where we need to go in order to improve and further develop future microarray analysis technology.

This book contains one introductory chapter and 19 technical chapters, dealing with specific methods or class of methods. As illustrated in Table 1, the technical chapters are roughly grouped into two broad categories, namely *data preparation* and *exploratory data analysis* respectively. Partitioning the chapters into these areas largely mirrors the current state of the art in the field. Different protocols, experimental conditions, analysis goals, data complexity, and sources of systematic variation in microarray experiments normally require different ways for selecting and preparing the raw data obtained from the detection devices. These methods range from missing value imputation and normalization to feature subset selection and data integration. Collectively, we refer to these methods as data preparation or *pre-processing* techniques. Once the final format for the data is achieved, data exploration or analysis can commence. This part of the data analysis process is referred to as *exploratory data analysis.* Typical exploratory analysis tasks include *classification* (or class prediction), *clustering* (or automatic classification), correlation and association analysis, and others.

As much as possible, the chapters are presented in an order that reflects the overall data analysis process. In Chapter 1, we provide an introduction to microarray analysis with the aim of (a) providing an easy-to-understand description of the entire process, and (b) establishing a common terminology. First, we recapitulate the biological and technological background of microarray hybridization experiments. This includes the main

types of arrays that exist, aspects of their protocols, and what kind of quantities they are measuring. Second, we categorize the classes of questions life scientists hope to answer with microarray experiments, and what kind of analytical tasks they imply. Third, we describe the entire process from the inception of a scientific question or hypothesis, to the design and execution of a microarray experiment, and finally to data preparation, analysis, and interpretation. We then discuss some of the conceptual and practical difficulties the experimenter faces when choosing and applying specific data analysis techniques. The remaining chapters are intended to shed more detailed light on these issues.

Table 1. Roadmap to the content of the book.

Part	Topic	Chapter#
	Introduction: Overview of microarray analysis process	1
Pre-processing	Foundations, issues, and methods	2
	Missing value imputation	3
	Error handling and normalization	4
	Singular value decomposition, principal component analysis	5
	Feature selection: established and recent techniques	6
Classification	Statistical foundations and methods	7
	Bayesian networks	8
	Support vector machines	9
	Weighted flexible compound covariate method and decision trees	10
	Artificial neural networks	11
	k-nearest neighbor and genetic algorithms	12
Cluster Analysis	Overview and review of some methods	13
	Hierarchical clustering methods	14
	Self-organizing maps	15
	Other non-hierarchical methods	16
Other	Correlation and association analysis methods	17
	Functional interpretation analytical results	18
Tools	Systematic review of free and commercial software	19
	Managing microarray data analysis: workflow and process	20

Chapter 2 addresses the issue of data pre-processing in microarray analysis in general. It is written for the newcomer to this field and explains the basic concepts and provides a useful vocabulary. It discusses the motivation for normalization, data centralization, data re-scaling, and missing value imputation. This chapter represents an introduction to the Chapters 2 to 5.

Chapter 3 presents three different methods for missing value imputation in microarray data. This includes a *k-nearest-neighbor* approach, a method

based on *singular value decomposition,* and *row averaging*. Practical guidelines are presented for using publicly available free software tools.

Chapter 4 discusses various sources of errors in microarray data, and then proceeds with a detailed discourse on normalization. In contrast to Chapter 2, the focuses is on mathematical considerations.

Chapter 5 is concerned with a major problem in microarray data analysis – the so-called *large-p-small-n problem* also known as the *curse of dimensionality.* This refers to the fact that for many microarray experiments the number of variables (genes) exceeds the number of observations (samples) by a factor of 10, 1000, or more. Feature selection and dimension reduction methods refer to techniques designed to deal with this "curse". The chapter discusses the use – and misuse – of *singular value decomposition* and *principal component analysis.*

Chapter 6 is a survey of several important feature selection techniques used to ward off the curse of dimensionality. First, it presents classic *filter* and *wrapper* approaches and some recent variants of explicit feature selection. Second, it outlines several feature weighting techniques including WINNOW and Bayesian feature selection. Third, towards the end, the chapter describes some recent work on feature selection for clustering tasks, a subject that has been largely neglected.

Chapter 7 discusses statistical issues arising in the classification of gene expression data. This chapter introduces the statistical foundations of classification. It provides an overview of traditional classifiers, such as linear discriminant analysis and nearest neighbor classifiers, in the context of microarray analyses. The general issues of feature selection and classifier performance assessment are discussed in detail.

Chapter 8 looks at Bayesian networks for the classification of microarray data. It introduces the basic concept of this approach, and reports on a study where the performance of Bayesian networks was compared with other state-of-the-art classifiers.

Chapter 9 describes a classification method that has been gaining increasing popularity in the microarray arena – *support vector machines* (SVMs). It provides an informal theoretical motivation of SVMs, both from a geometric and algorithmic perspective. Instead of focusing on mathematical completeness, the intention of this chapter is to provide the practitioner with some "rules of thumb" for using SVMs in the context of microarray data. Finally, pointers to relevant, publicly available free software resources are given.

Chapter 10 reports on a recent case study of gene expression analysis in lung cancer. The authors describe the *weighted flexible compound covariate* method for classifying the microarray data. They also demonstrate how this relatively new method is related to decision trees.

Chapter 11 deals with a widely used machine learning technique called *artificial neural networks* (ANNs). The authors describe the application of ANNs to micrarray classification task. They discuss how a principal component analysis, *cross-validation* and *random permutation tests* can be employed to improve and evaluate the predictive performance of ANNs. The problem of extracting important genes from a constructed ANN is also addressed.

Chapter 12 represents the last chapter on classification. It presents the *k-nearest-neighbor* strategy and *genetic algorithms* for classifying microarray data. It discusses the general motivation and the concepts of these methods, and demonstrates their performance on microarray data sets. The authors provide references to publicly available free software resources.

Chapter 13 provides an overview of the major types of clustering problems and techniques for microarray data. It focuses on crucial design and analytical aspects of the clustering process. The authors provide some important criteria for selecting clustering methods. Furthermore, the chapter describes a scheme for evaluating clustering results based on their relevance and validity (both computational and biological).

Chapter 14 addresses hierarchical clustering methods in the context of microarray data. The discussed methods include hierarchical clustering methods, including *adaptive single linkage clustering,* a new method designed to provide adaptive cluster detection while maintaining scalability. Furthermore, the chapter provides examples using both simulated and real data.

Chapter 15 presents *self-organizing maps* (SOMs) for clustering microarray data. It discusses question such as: How do these models work? Which are their advantages and limitations? Which are the alternatives? In answering these questions, this chapter constitutes a rich source of practical guidelines for using SOMs to analyze microarray data.

Chapter 16 examines a number of non-hierarchical clustering algorithms for microarray analysis, namely *cluster affinity search technique,* the famous *k-means* technique, *partitioning around medoids,* and *model-based clustering*. The chapter puts emphasis on the practical aspects of these algorithms, such as guidelines for parameter setting, the specific algorithmic properties, and practical tips for implementation.

Chapter 17 addresses correlation and association analysis methods. It addresses questions that should help the user to assess the limitations and merits of these methods, such as: How to statistically measure the strength between two variables and test their significance? What is correlation, what is association? Which conclusions do correlation and association analysis allow in the context of microarray data?

Chapter 18 discusses the global functional interpretation of gene expression experiments. After a researcher has found differentially expressed

genes using one of the above described methods, he must face the challenge of translating his results into a better understanding of the underlying biological phenomena. This chapter shows how this can be achieved.

Chapter 19 provides an overview of both publicly available, free software and commercial software packages for analyzing microarray data. The aim of this review is to provide an overview of various microarray software categorized by their function and characteristics. This review should be a great help for those who are currently consider obtaining such software.

Finally, Chapter 20 describes the microarray data analysis from process perspective, highlighting practical issues such as project management and workflow considerations.

The book is designed to be used by the practicing professional tasked with the design and analysis of microarray experiments or as a text for a senior undergraduate- or graduate level course in analytical genetics, biology, bioinformatics, computational biology, statistics and data mining, or applied computer science. In a quarter-length course, one lecture can be spent on each chapter, 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 more of the formal background of the discussed methods. Each chapter includes recommendations for further reading. Questions or comments about the book should be directed to the editors by e-mail under *dp.berrar@ulster.ac.uk, w.dubitzky@ulster.ac.uk,* or *granzow@4T2consulting.de.* For further details on the editors, please refer to the following URL:

http://www.infj.ulst.ac.uk/~cbbg23/interests.html.

Daniel Berrar

Werner Dubitzky

Martin Granzow