Preface

Although the industry once suffered from a lack of qualified targets and candidate drugs, lead scientists must now decide where to start amidst the overload of biological data. In our opinion, this phenomenon has shifted the bottleneck in drug discovery from data collection to data analysis, interpretation and integration.

—Life Science Informatics, UBS Warburg Market Report, 2001

One of the most promising tools available today to researchers in life sciences is the microarray technology. Typically, one DNA array will provide hundreds or thousands of gene expression values. However, the immense potential of this technology can only be realized if many such experiments are done. In order to understand the biological phenomena, expression levels need to be compared between species or between healthy and ill individuals or at different time points for the same individual or population of individuals. This approach is currently generating an immense quantity of data. Buried under this humongous pile of numbers lays invaluable biological information. The keys to understanding phenomena from fetal development to cancer may be found in these numbers. Clearly, powerful analysis techniques and algorithms are essential tools in mining these data. However, the computer scientist or statistician that does have the expertise to use advanced analysis techniques usually lacks the biological knowledge necessary to understand even the simplest biological phenomena. At the same time, the scientist having the right background to formulate and test biological hypotheses may feel a little uncomfortable when it comes to analyzing the data thus generated. This is because the data analysis task often requires a good understanding of a number of different algorithms and techniques and most people usually associate such an understanding with a background in mathematics, computer science, or statistics.

Because of the huge amount of interest around the microarray technology, there are quite a few books available on this topic. Many of the few available texts concentrate more on the wet lab techniques than on the data analysis aspects. There are several books that review the topic in a somewhat superficial manner, covering everything there is to know about data analysis of microarrays in a couple of hundred pages or less. Other available books focus on excruciating details that only developers of analysis packages find useful. Others are simple proceedings of conferences, gathering together unrelated

papers that focus on very specific aspects and topics. Overall, I felt there was a need for a good, middle-of-the-road book that would cover topics in sufficient details to make it possible for readers to really understand what is going on, but without overwhelming details or intimidating heavy formalisms or notations.

At the same time, the R environment has started to dominate heavily everything that is done in terms of data analysis in this area. Again, while many good books on R as a programming and analysis language are available, I felt that a book that would allow the reader to become competent in the analysis of microarray data by providing: i) everything needed to learn the basics of R, ii) the basics of the microarray technology, as well as iii) the understanding necessary in order to apply the right tools to the right problems would be beneficial.

Audience and prerequisites

The goal of this book is to fulfill this need by presenting the main computational techniques available in a way that is useful to both life scientists and analytical scientists. The book tries to demolish the imaginary concrete wall that separates biology and medicine from computer science and statistics and allow the biologist to be a refined user of the available techniques, as well as be able to communicate effectively with computer scientists and statisticians designing new analysis techniques. The intended audience includes as a central figure the researcher or practitioner with a background in the life sciences that needs to use computational tools in order to analyze data. At the same time, the book is intended for the computer scientists or statisticians who would like to use their background in order to solve problems from biology and medicine. The book explains the nature of the specific challenges that such problems pose as well as various adaptations that classical algorithms need to undergo in order to provide good results in this particular field.

Finally, it is anticipated that there will be a shift from the classical compartmented education to a highly interdisciplinary approach that will form people with skills across a range of disciplines crossing the borders between traditionally unrelated fields, such as medicine or biology and statistics or computer science. This book can be used as a textbook for a senior undergraduate or graduate course in such an interdisciplinary curriculum. The book is suitable for a data analysis and data mining course for students with a background in biology, molecular biology, chemistry, genetics, computer science, statistics, mathematics, etc.

Useful prerequisites for a biologist include elementary calculus and algebra. However, the material is designed to be useful even for readers with a shaky mathematical foundation since those elements that are crucial for the topic are fully discussed. Useful prerequisites for a computer scientist or mathematician include some elements of genetics and molecular biology. Once

again, such knowledge is useful but not required since the essential aspects of the technology are covered in the book.

Aims and contents

The first and foremost aim of this book is **to provide a clear and rigorous description of the algorithms without overwhelming the reader with the usual cryptic notation or with too much mathematical detail.** The presentation level is appropriate for a scientist with a background in life sciences. Little or no mathematical training is needed in order to understand the material presented here. Those few mathematical and statistical facts that are really needed in order to understand the techniques are completely explained in the book at a level that is fully accessible to the non-mathematically minded reader. The goal here was to keep the level as accessible as possible. The mathematical apparatus was voluntarily limited to the very basics. The most complicated mathematical symbol throughout the book is the sum of n terms: $\sum_{i=1}^{n} x_i$. In order to do this, certain compromises had to be made. The definitions of many statistical concepts are not as comprehensive as they could be. In certain places, giving the user a powerful intuition and a good understanding of the concept took precedence over the exact, but more difficult to understand, formalism. This was also done for the molecular biology aspects. Certain cellular phenomena have been presented in a simplified version, leaving out many complex phenomena that we considered not to be absolutely necessary in order to understand the big picture.

A second specific aim of the book is **to allow a reader to learn the R environment and programming language using a hands-on and example-rich approach.** From this perspective, the book should be equally useful to a large variety of readers with very different backgrounds. No previous programming experience is required or expected from the reader. The book includes chapters that describe everything from the basic R commands and syntax to rather sophisticated procedures for quality control, normalization, data analysis and machine learning. Everything from the simplest commands to the most complex procedures is illustrated with R code. All analysis results shown in the book are actual results produced by the code shown in the text and therefore, all code shown is free from spelling or syntax errors.

A third specific aim of the book is **to allow a microarray user to be in a position to make an informed choice as to what data analysis technique to use in a given situation, even if using other analysis packages.** The existing software packages usually include a very large number of techniques, which in turn use an even larger number of parameters. Thus, the biologist trying to analyze DNA microarray data is confronted with an overwhelming number of possibilities. Such flexibility is absolutely crucial because each data set is different and has specific particularities that must be taken into account when selecting algorithms. For example, data sets obtained in different laboratories have different characteristics, so the choice of normal-

ization procedures is very important. However, such wealth of choices can be overwhelming for the life scientist who, in most cases, is not very familiar with all intricacies of data analysis and ends up by always using the default choices. This book is designed to help such a scientist by emphasizing at a high level of abstraction the characteristics of various techniques in a biological context.

As a text designed to bridge the gap between several disciplines, the book includes chapters that would give all the necessary information to readers with a variety of backgrounds. The book is divided into two parts. The first part is designed to offer an overview of microarrays and to create a solid foundation by presenting the elements from statistics that constitute the building blocks of any data analysis. The second part introduces the reader to the details of the techniques most commonly used in the analysis of microarray data.

Chapter 2 presents a short primer on the central dogma of molecular biology and why microarrays are useful. This chapter is aimed mostly at analytical scientists with no background in life sciences. **Chapter 3** briefly presents the microarray technology. For the computer scientist or statistician, this constitutes a microarray primer. For the microarray user, this will offer a bird's-eye view perspective on several techniques emphasizing common as well as technology-specific issues related to data analysis. This is useful since many times the users of a specific technology are so engulfed in the minute details of that technology that they might not see the forest for the trees.

Chapter 4 discusses a number of important issues related to the reliability and reproducibility of microarray data. This discussion is important both for the life scientist who needs to understand the limitations of the technology used, as well as for the computer scientist or statistician who needs to understand the intrinsic level of noise present in this type of data.

Chapter 5 constitutes a short primer on digital imaging and image processing. This chapter is mostly aimed at the life scientists or statisticians who are not familiar with digital image processing.

Chapter 6 is an introduction to the R programming language. This chapter discusses basic concepts from the installation of the R environment to the basic syntax and concepts of R.

Chapter 7 presents the Bioconductor project and briefly illustrates its capabilities with some simple examples. This chapter was contributed by one of the founders of the Bioconductor project, Vincent Carey, currently an Associate Professor of Medicine (Biostatistics) at Harvard Medical School, and an Associate Biostatistician in the Department of Medicine at the Brigham and Women's Hospital in Boston.

Chapters 8, **9**, and **11** focus on some elementary statistics notions. These chapters will provide the biologist with a general perspective on issues very intimately related to microarrays. The purpose here is to give only as much information as needed in order to be able to make an informed choice during the subsequent data analysis. The aim of the discussion here is to put things in the perspective of somebody who analyzes microarray data rather than offer a full treatment of the respective statistical notions and techniques. **Chapter 9**

discusses several important distributions, **Chapter 11** discusses the classical hypothesis testing approach, and **Chapter 12** applies it to microarray data analysis. **Chapter 10** uses R to illustrated the basic statistical tools available in R for descriptive statistics and basic built-in distributions.

Chapter 13 presents the family of ANalysis Of VAriance methods intensively used by many researchers to analyze microarray data. **Chapter 14** discusses the more general linear models and illustrates them in R. **Chapter 15** uses some of the ANOVA and linear model approaches in the discussion of various techniques for experiment design.

Chapter 16 discusses several issues related to the fact that microarrays interrogate a very large number of genes simultaneously and its consequences regarding data analysis.

Chapters 17 and **18** present the most widely used tools for microarray data analysis. In most cases, the techniques are presented using real data. Chapter 17 includes several techniques used in exploratory analysis, when there is no known information about the problem and the task is to identify relevant phenomena as well as the parameters (genes) that control them. The main techniques discussed here include box plots, histograms, scatter plots, volcano plots, time series, principal component analysis (PCA), and independent component analysis (ICA). The clustering techniques described in Chapter 18 include K-means, hierarchical clustering, biclustering, partitioning-around-medoids, and self-organizing feature maps. Again, the purpose here is to explain the techniques in an unsophisticated yet rigorous manner. The all-important issue of when to use a specific technique is discussed on various examples emphasizing the strengths and weaknesses of each individual technique.

Chapter 19 discusses specific quality control issues characteristic to Affymetrix and Illumina data. These are illustrated using R functions and packages applied on real data sets. Tools such as intensity distributions, box plots, RNA degradation curves, and quality control metrics are used to illustrate problems ranging from array saturation, to RNA degradation, and annotation issues. Plots illustrating various problems are shown side-by-side with plots showing clean data such that the reader can understand and learn what to look for in such plots.

Chapter 20 concentrates on data preparation issues. Although such issues are crucial for the final results of the data mining process, they are often ignored. Issues such as color swapping, color normalization, background correction, thresholding, mean normalization, etc., are discussed in detail. This chapter will be extremely useful both to the biologist, who will become aware of the different numerical aspects of the various preprocessing techniques, and to the computer scientist, who will gain a deeper understanding of various biological aspects, motivations, and meanings behind such preprocessing. Again, all normalization issues are illustrated using R functions and packages applied on real data sets.

Chapter 21 presents several methods used to select differentially regulated genes in comparative experiments.

Chapter 22 discusses the Gene Ontology, including its goal, structure, annotations, and some statistics about the data currently available in it. **Chapter 23** shows how GO can be used to translate lists of differentially expressed genes into a better understanding of the underlying biological phenomena. Just when you think this is easy, **Chapter 24** comes to tell you about the many mistakes and issues that could appear in this GO profiling. **Chapter 25** reviews more than a dozen tools that are currently available for this type of functional analysis.

Chapter 26 somehow reverses the direction considering the problem of how to select the microarrays that are best suited for investigating a given biological hypothesis.

Chapter 27 discusses some of the problems that can be caused by the fact that the same biological entity may have different IDs in different public databases. Some tools that allow a mapping from one type of ID to another are discussed and compared.

Chapter 28 takes the analysis to the next level, using a systems biology approach that aims to take into consideration the way genes are known to interact with each other. This type of knowledge is captured in collections of signaling pathways available from various sources. This chapter discusses various approaches currently available for the analysis of signaling pathways.

Chapter 29 is a brief review of several machine learning techniques that are widely used with microarray data. Since unsupervised methods are discussed in **Chapter 18**, this chapter focuses on supervised methods including linear discriminants, feed-forward neural networks, and support vector machines.

Finally, the last chapter of the book presents some conclusions as well as a brief presentation of some novel techniques expected to have a great impact on this field in the near future.

Road map

This book can be used in several ways, depending on the background of the reader and the goals pursued. The chapters can be combined in various ways, allowing an instructor to tailor a course to the specific background and expectations of a given audience.

Some of the courses (with or without a laboratory component) that can be easily taught using this book include:

1. Introduction to statistics: Chapters 8, 9, 11, 12, 13, 14, 15, 16
2. Introduction to R and Bioconductor: Chapters 6, 7, 10, 14, 17, 18, 29

3. Microarray data analysis for life scientists: Chapters 3 – 30
4. Microarray data analysis for computer scientists (including R and Bioconductor): Chapter 2 – 30
5. Quality control and normalization techniques: Chapters 19, 20
6. Interpretation of high-throughput data: GO profiling and pathway analysis: Chapters 22 – 28

This book focuses on R and Bioconductor. The reader is advised to install the software and actually use it to perform the analysis steps discussed in the book. The accompanying CD includes all code used throughout the book.

Acknowledgments

The author of this book has been supported by the Biological Databases Program of the National Science Foundation under grants number NSF-0965741 and NSF-0234806, the Bioinformatics Cell, MRMC US Army – DAMD17-03-2-0035, National Institutes of Health – grant numbers 1R01DK089167-01, R01-NS045207-01 and R21-EB000990-01, and Michigan Life Sciences Corridor grant number MLSC-27. Many thanks to Sylvia Spengler, Director of the Biological Databases program, National Science Foundation, Salvatore Sechi, Program Manager NIDDKD, Peter McCartney, Program Director, Division of Biological Infrastructure, National Science Foundation, Peter Lyster, Program Director in the Center for Bioinformatics and Computational Biology, NIGMS without whose support our research and this book would not have been possible.

I would like to express my gratitude to Dr. Robert Romero, Chief of the Perinatology Research Branch of the National Institute of Child Health and Human Development, who has been a tremendous source of inspiration and a role model in my research. I am also very grateful for his strong support which allowed me to dedicate more time to my research and to other scholarly endeavors such as writing this book.

My deepest gratitude goes to Robert J. Sokol, M.D., The John M. Malone, Jr., M.D., Endowed Chair, and Director of the C.S. Mott Center for Human Growth and Development, Distinguished Professor of Obstetrics and Gynecology in the Wayne State University School of Medicine. His efforts and support were the main reasons that kept me at Wayne State when I had great opportunities elsewhere. Also, his generosity in creating the endowed chair in systems biology that bears his name – and that I am currently holding – made possible the creation of the strong research group of which I am a part.

My colleague Adi Laurentiu Tarca had a substantial contribution to this

book. This consisted of a first draft of the linear models chapter, as well as numerous code snippets used throughout several other chapters to illustrate various concepts and techniques. He also contributed with a very thorough critical reading of many other chapters followed by very useful comments and suggestions.

My thanks also go to Vincent Carey, Associate Professor of Medicine (Biostatistics) at Harvard Medical School, and an Associate Biostatistician in the Department of Medicine at the Brigham and Women's Hospital in Boston. Vincent is one of the founders of the Bioconductor project and contributed Chapter 7 to this book. He was also very helpful in answering several questions regarding Bioconductor and various packages used throughout this book.

My Ph.D. students Calin Voichita, Michele Donato, Cristina Mitrea, Dorina Twigg, Munir Islam, Rebecca Tagett, and my visiting postdoctoral researcher Josep Maria Mercader (currently back at his home institution – Barcelona Supercomputer Center) contributed enormously to this book, first by accepting to be my guinea pigs for the examples and explanations used in the book, and then by proof-reading all 1,081 pages of this book – several times. In spite of their tremendous efforts, typos and small errors probably continue to exist here and there. If this is the case, the fault is entirely mine. I am sure that for each such surviving typo, somewhere in my inbox there is an email from one of them, bringing it to my attention.

My research associate, Zhonghui Xu, had a substantial contribution to this book by writing various R routines and snippets of code, in particular most of those in the chapters on normalization and quality control.

Gary Chase kindly reviewed the chapters on statistics (for the first incarnation of this book) and did so on incredibly short deadlines that allowed the book to be published on time. His comments were absolutely invaluable and made the book stronger than I initially conceived it.

I owe a lot to my colleague, friend and mentor, Michael A. Tainsky. Michael taught me everything I know about molecular biology and I continue to be amazed by his profound grasp of some difficult data analysis concepts. I have learned a lot during our collaboration over the past 10 years. Michael also introduced me to Judy Abrams who provided a lot of encouragement and has been a wonderful source of fresh and exciting ideas, many of which ended up in grant proposals. My colleague and friend Steve Krawetz was the first one to identify the need for something like Onto-Express (Chapter 23). During my leave of absence, he continued to work with my students and participated actively in the creation of our first Onto-Express prototype. Our collaboration over the past few years produced a number of great ideas and a few papers. Thanks also go to Jeff Loeb for our productive interaction as well as for our numerous discussions on corrections for multiple experiments.

Many thanks to Otto Muzik for sticking to the belief that our collaboration is worthwhile through the first, less than productive, year and for being such a wonderful colleague and friend. In spite of some personal differences, Otto did not hesitate to offer his help when my father was diagnosed with cancer.

I am very grateful to him for his offer to help, even though my father passed away before he could take advantage of Otto's kind offer.

Jim Granneman and Bob Mackenzie knocked one day on my door looking for some help analyzing their microarray data. One year later, we were awarded a grant of over 3.5 million from the Michigan Life Science Corridor. This allowed me to shift some of my effort to this book. Currently, Jim and Bob are my coPIs on a 1.5 million NIH grant which is funding some very exciting research. Thank you all for involving me in your work.

I am grateful to Soheil Shams, Bruce Hoff, Anton Petrov and everybody else with whom I interacted during my one year sabbatical at BioDiscovery. Bruce, Anton, and I worked together at the development of the ANOVA based noise sampling method described in Chapter 21 based on some initial experiments by Xiaoman Li. Figures 18.6, 18.12, and 18.15 were initially drawn by Bruce.

During my stay in Los Angeles, I met some top scientists, including Michael Waterman and Wing Wong. I learned a lot from our discussions and I am very grateful for that. Mike Waterman provided very useful comments and suggestions on the noise sampling method in Chapter 21 and later, on OntoExpress described in Chapter 23. Wing Wong and his students at UCLA and Harvard have done a great job with their software dChip, which is a great first stop every time one has to normalize Affymetrix data.

Gary Churchill provided useful comments on the work related to the ANOVA based noise sampling technique presented in Chapter 21. The discussions with him and Kathy Kerr helped us a lot. Also, their ANOVA software for microarray data analysis represent a great tool for the research community. We used some of their software to generate the images illustrating the LOWESS transform in Chapter 20.

John Quackenbush kindly provided a wonderful sample data set that we used in the PCA examples in Chapter 17. He and his colleagues at TIGR designed, implemented and made available to the community a series of very nice tools for microarray data analysis including the Multiple Experiment Viewer (MEV) used to generate some images for Chapter 17.

About the author

Sorin Drăghici has obtained his B.Sc. and M.Sc. degrees in Computer Engineering from "Politehnica" University in Bucharest, Romania followed by a Ph.D. degree in Computer Science from University of St. Andrews, United Kingdom (third oldest university in UK after Oxford and Cambridge). Besides this book, he has published two other books, several book chapters, and over 100 peer-reviewed journal and conference papers. He is inventor or co-inventor on several patent applications related to biotechnology, high-throughput meth-

ods and systems biology. He is currently an editor of IEEE/ACM Transactions on Computational Biology and Bioinformatics, the Journal of Biomedicine and Biotechnology, and the International Journal of Functional Informatics and Personalized Medicine. He is also active as a journal reviewer for 15 international technical journals related to bioinformatics and computational biology, as well as a regular NSF and NIH panelist on biotechnology and bioinformatics topics.

Currently, Sorin Drăghici is holding the Robert J. Sokol, MD Endowed Chair in Systems Biology in the Department of Obstetrics and Gynecology in the School of Medicine, as well as joint appointments as Professor in the Department of Clinical and Translational Science, School of Medicine, and the Department of Computer Science, College of Engineering, Wayne State University. He is also the Chief of the Bioinformatics and Data Analysis Section, Perinatology Research Branch, National Institute for Child Health and Development, NIH, as well as the head of the Intelligent Systems and Bioinformatics Laboratory (ISBL, http://vortex.cs.wayne.edu).

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, National Institutes of Health, any other funding agency, nor those of any of the people mentioned above.