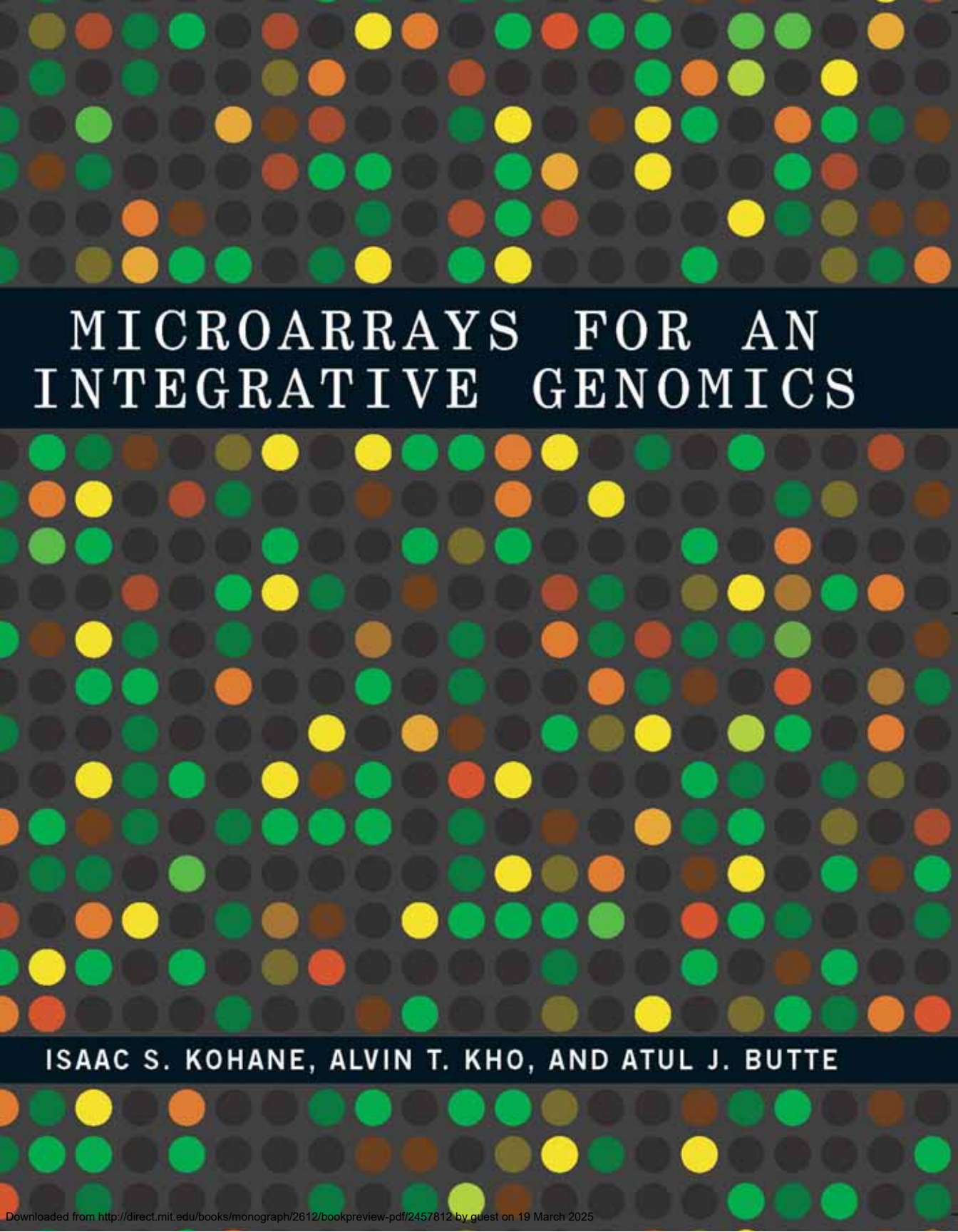

Microarrays for an Integrative Genomics

Computational Molecular Biology
Sorin Istrail, Pavel Pevzner, and Michael Waterman, editors

Computational molecular biology is a new discipline, bringing together computational, statistical, experimental and technological methods, which is energizing and dramatically accelerating the discovery of new technologies and tools for molecular biology. The MIT Press Series on Computational Molecular Biology is intended to provide a unique and effective venue for the rapid publication of monographi, textbooks, edited collections, reference works, and lectures notes of the highest quality.

Computational Molecular Biology: An Algorithmic Approach
Pavel Pevzner, 2000

Computational Modeling of Genetic and Biochemical Networks
edited by James Bower and Hamid Bolouri, 2001

Current Topics in Computational Molecular Biology
Tao Jiang, Ying Xu, and Michael Q. Zhang, editors, 2002

Gene Regulation and Metabolism: Postgenomic Computation Approaches
Julio Collado-Vides, editor, 2002

Microarrays for an Integrative Genomics
Isaac S. Kohane, Alvin T. Kho, and Atul J. Butte, 2003

Microarrays for an Integrative Genomics

Isaac S. Kohane, Alvin T. Kho, and Atul J. Butte

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in in Computer Modern by the authors using the LaTeX typesetting system and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Kohane, Isaac S.
Microarrays for an integrative genomics / Isaac S. Kohane, Alvin Kho, Atul J. Butte
p. cm.—(Computational molecular biology)
Includes bibliographical references.
ISBN 0-262-11271-X (hc.: alk. paper)
1. DNA microarrays. 2. Gene expression—Analysis—Automation. 3. Bioinformatics. I. Kho, Alvin. II. Butte, Atul J. III. Title. IV. Series.
QP624.5.D726 K686 2002
572.8′6–dc21 2002022663

Contents

Foreword

The impact of microarray measurements on biology and bioinformatics has been astounding. Starting from virtually no literature a few years ago, this field has come to dominate many conferences and journals. As an example, the Intelligent Systems for Molecular Biology conference, the annual meeting of the International Society for Computational Biology held in Copenhagen in 2002, had almost 50% of its papers in the areas addressed directly or indirectly within this book. Four years ago, there were none. Bioinformatics has always been driven by the availability of data—sequential, structural, and most recently functional. The availability of sequence data brought into biology a cadre of computer scientists with special skills in string processing. The availability of structural data brought in technical experts in visualization and computational geometry. This most recent development—the availability of relatively large data sets measuring the expression of genes within cells—has helped attract yet another group of scientists—the data miners, machine learners and statisticians.

In many ways, the impact of this data on biology and informatics can be summarized in figure 1.4 of this book—the world has not quite been turned upside down, but it certainly has been turned on its side! A decade ago, if confronted with the data matrix shown on the right of this figure, a well-trained information scientist would say "This is ridiculous. Why would you ask me to analyze a data set where you clearly have a profoundly under-determined problem? There's not enough data here to distinguish between any of the zillion hypotheses that could be consistent with this data set. And who designed these experiments, anyway? How can you measure so many features of such a few examples?" Yet, these experiments are proceeding and are making major contributions to our understanding of how gene systems interact, how to distinguish different types of cancer, and how to measure the impact of the environment on a cell. Our information scientist friend is, in some sense, correct about the relative paucity of data. (Have you ever tried to convinced a biologist holding a microarray with 45,000 spots that this is a relatively data-poor exercise? It's not fun.) However, the information scientist has missed the point about the design and analysis of these experiments. These data sets do indeed contain gold, but the experiments (as for all experiments) must be considered carefully in both the design and implementation phases in order to maximize value. This is where the authors of this book have made a contribution. They start from the premise that these experiments offer great potential, but must be performed and analyzed carefully. They set the context of traditional reductionist biology, and then go on to discuss the design, analysis, storage and interpretation of this first generation of functional genomics experiments. The writing is lively and candid,

and the examples are taken from an array of applications. The authors' practical experience in dealing with this data comes through, and they intersperse practical advice with philosophical reverie. Sometimes, these two merge into important discussions such as on the role of ontologies in making sense of these data sets, or on the challenges of linking microarray results with phenotypic data pertinent to human disease.

The functional genomics revolution is here. We do not know how it will change our view of biology and medicine. They are both much more likely to become quantitative and systematic (as opposed to qualitative and reductionist). The informatics techniques required to address this revolution are not entirely clear, but this text gets us started in the right direction.

Russ Altman
Stanford University
March 2002

Preface

Three years ago, when a colleague would approach us with questions about functional genomics and the informatics techniques required to leverage the data obtained from measurement techniques such as DNA microarrays, we had a standard response: "Come listen to a 1-hour presentation by one of us and you'll have a foundation for further discussions." Since then, this response has become inadequate. First, the field can hardly be summarized in even eight 1-hour lectures, and second, the growth in the number of potential collaborators has far outstripped the time available to us to make the necessary presentations.

In early 2000, Ben Reis, one of the graduates of the Children's Hospital Informatics Program, had the inspired idea of formalizing our introductory presentations into a book. We immediately agreed that this was a timely suggestion and to its credit, so did The MIT Press. In our teaching duties in several courses within the Division of Health Sciences and Technology (HST) at Harvard/MIT the range of topics within functional genomics that we were covering in formal presentations grew rapidly. Subsequently, with the inception of the Bioinformatics and Integrative Genomics training program at HST and the development of a Genomic Medicine course at HST, we felt the need for this book all the more acutely.

Organization of the text

We recognize that the readership of this book will be varied due to the intrinsically multidisciplinary nature of the functional genomics enterprise (as will be emphasized in the introductory chapter). Accordingly we outline the content of the following chapters so that readers may choose for themselves the path that suits them. Nonetheless, our intent and contention is that the current ordering of the chapters provides the most efficient way of acquiring the content of this book.

Introduction. Here we establish the motivation and the scope of this book and touch upon substantial obstacles to success in the successful application of bioinformatics to an integrative genomics. The notion of an interdisciplinary functional genomics pipeline is introduced. We also review which kind of readers might find this book worthwhile. The promise and limitations of functional genomics techniques, the nature of various kinds of genomic data, and the central role played by the discipline of bioinformatics are outlined. For those who have a limited background in biological sciences, there is a subsection on the basic minimum of molecular biology concepts that will be needed to grasp the the following chapters.

Chapter 1. Experimental Design. This chapter develops a framework for ap-

proaching the design of microarray-driven functional genomics experiments. Very little here is quantitative or mathematical. Rather the emphasis is on ways of thinking about the design of experiments and how it might impact the yield of these experiments. We address challenges that are particular to computer scientists (*e.g.*, defining a figure merit for the performance of the bioinformatics algorithms) and to biologists (*e.g.*, discarding potentially valuable data using formal decision theory because of the scale issues in massively parallel data acquisition using noisy measurement devices), respectively. In exploring the design issues we introduce the functional genomics clustering dogma, the broad machine-learning categories of supervised and unsupervised learning, and the nature of the analyses developed using these techniques.

Chapter 2. Microarray Measurements to Analyses. We lay the foundations for performing analyses of microarray data sets. This is the first of the more quantitative and mathematical chapters. We start with a discussion of the acquisition of digital data from the two most widely employed classes of microarrays. Then we consider the two most generic problems of comparing gene expression within a single microarray, *i.e.*, *intra*-array analyses, and comparing expression across microarrays, *i.e.*, *inter*-array analyses; in so doing, we introduce the fundamental concept of (dis)similarity and similarity measures and the several kinds of such measures. These measures become the building blocks for the genomic data-mining techniques described in the following chapter.

Chapter 3. Genomic Data-Mining Techniques. When gene expression is measured in more than two samples, gene expression patterns have to be analyzed using methods that consider the coordinated interactions of genes across multiple conditions. This chapter assesses the components of biomedical experiments that can be included in a data-mining investigation. We then cover the most commonly used analytic techniques, discussing the advantages and disadvantages of each technique, as well as the postanalysis process. Where appropriate, we provide pseudocode that will allow readers with some training in computer science to understand the details of the most often used and cited data-mining algorithms. The emerging field of genetic network reverse engineering is also introduced here.

Chapter 4. Bio-Ontologies, Data Models, Nomenclature. This chapter addresses possibly the least exciting but the most pressing bioinformatics need for genomic research: creating and using comprehensive annotations of gene function, storing and organizing microarray expression data, and ensuring standardized access to these data. We review current efforts to create formalized systems of description of gene function and the various kinds of "ontologies" that support these descriptions. The challenge to design "standardized data models" for the storage of microarray data

is addressed and the principal contenders claiming to be this standard are reviewed. Naming schemes—nomenclatures—most applicable to gene expression studies are described. Nomenclatures, data models, and ontologies are placed in a perspective of the general problem of analyzing the results of functional genomics experiments. Tools that leverage these standardization efforts and the on-line published literature are also described.

Chapter 5. From Functional Genomics to Clinical Relevance: Getting the Phenotype Right. Here we address the process of translating the functional genomics research agenda into one of clinical relevance. We place in this perspective the value and deficiencies of electronic medical records and standardized clinical vocabularies. Although by no means comprehensive, we provide the highlights of the privacy issues (*e.g.*, the implications of the Health Insurance Portability and Accountability Act, anonymization, cryptographic identifiers, *etc.*) that are most likely to have an impact on the clinical application of genomic technologies.

Chapter 6. The Near Future. As the techniques and goals of functional genomics are in rapid flux, we engage in some short-term forecasting to guide readers planning in this time window. Microarray technologies being developed and recently released are previewed. In this context, the problem of comparing expression measurements across generations of microarray measurement platforms is appraised. More broadly, the different kinds of software required for the successful functional genomics enterprise are described. Finally, a model to meet the training needs of this new discipline is outlined.

Acknowledgments

ISK would like to thank his two co-authors: Without their belief in the worth of this collaborative enterprise and without their expertise, this book would never have been completed. Dozens of colleagues generously provided of their knowledge and resources including Hamish Fraser, Steven Greenberg, Winston Kuo, Ashish Nimgaonkar, Peter Park, Marco Ramoni, Alberto Riva, Ben Reis, Zoltan Szallasi, Peter Szolovits, and Christine Tsien. Our colleagues in industry, notably Bill Craumer, John Hart, and Bill Buffington were equally generous. Of course, all errors and misinterpretations remain the authors'. David Ruckle is due thanks for his effective transmogrification of the authors' scribbles into attractive illustrations. To Marie Boyle, ISK offers his gratitude for her unparalleled organization skills. Robert Prior would do any publisher proud for his thoughtfulness and integrity. Elaine and Leo provided a quiet and hospitable haven for ISK before some particularly tight deadlines. To Heidi, ISK is thankful for the love, understanding and encouragement that enabled this work to proceed even when sunny days, and other enticing distractions beckoned. Finally, to Judith and the late Akiva Kohane who transmitted the core values of which this effort is a small reflection, ISK will always be in their debt.

ISK's time on this project has been funded in part through the generosity of the John F. and Virginia B. Taplin Award of the Harvard-MIT Division of Health Sciences and Technology as well through funding by the NIH including N01 LM-9-3536 "Personal Internetworked Notary and Guardian" from the National Library of medicine, HL066582-01 and HL-99-24 through the Program for Genomic Applications of the National Heart Lung and Blood Institute, U24 DK058739 "NIDDK Biotechnology Center" by the National Institute of Diabetes, Digestive and Kidney Diseases, 1R21 NS41764-01 "Functional Genomic Analysis of the Developing Cerebellum" by the National Institute of Neurological Disorders and Stroke, U01 CA091429-01 "Shared Pathology Informatics Network" of the National Cancer Institute, and P01 NS 40828-01 "Gene Expression in Normal and Diseased Muscle During Development" by the National Institute of Neurological Disorders and Stroke.

ATK is deeply indebted to his co-authors for their immense patience and learning in answering and enduring his oftentimes naïve biologic queries; To them and Marie Boyle, our work-besieged CHIP administrator, for their unrelenting cheerfulness and easy camaraderie during our writing spells. ATK is very grateful to David Rowitch for funding support via the National Institutes of Health grant 1R21 NS041764-01 "Functional Genomic Analysis of the Developing Cerebellum" throughout the duration of this project. Lastly, but not in the least, he expresses his deep gratitude to his parents Kwang Khoon and Rose Kho for their ever constant

love and affection.

AJB is first indebted to his co-authors for the stimulating discussions needed to prepare this book, many of which have resulted in the design and development of novel bioinformatics methods while this book was being written. In addition, without Marie Boyle, this book would never have been written or organized. AJB wishes to thank his friend and mentor, Dr. Isaac Kohane, for accepting him into the Children's Hospital Informatics Program four years ago. AJB sincerely owes his being in this field of research to Dr. Kohane's early vision and strong collaborative connections to biologists. AJB appreciates Dr. Kohane's continued support and development of his career in bioinformatics. AJB also wishes to thank his division chief, Dr. Joseph Majzoub, for his support of AJB's clinical and research environment, and Professor Peter Szolovits for his guidance and advice. AJB remembers his uncle, the late Dr. Prakash Kulkarni, who inspired him to become a pediatrician, and also thanks Dr. Simeon Taylor and Dr. Michael Quon, both of whom taught AJB everything he knows about the laboratory study of molecular biology and inspired him to enter the field of endocrinology. AJB wishes to thank his brother, Dr. Manish Butte, for being a role model with his academic pursuits and his discussions on the nature of biophysics and bioinformatics, and his parents Janardhan and Mangala Butte, for their love, support, and wisdom in exposing their children to computers at an early age. Finally, AJB wishes to thank his wife, Dr. Tarangini Deshpande, for her love, support, and her encouragement to complete this work.

During the writing of this work, AJB has been funded by and wishes to thank the Endocrine Fellows Foundation, the Genentech Center for Clinical Research and Education, the Lawson Wilkins Pediatric Endocrinology Society, the Harvard Center for Neurodegenerative Research, and the Merck-Massachusetts Institute of Technology partnership. AJB was also supported in part by the grant "Genomics of Cardiovascular Development, Adaptation, and Remodeling" funded by the National Heart Lung and Blood Institute's Program in Genomic Applications, U01 HL066582, the grant "Harvard-MIT-NEMC Research Training in Health Informatics", funded by the National Library of Medicine, 5T15 LM07092, and the grant "NIDDK Biotechnology Center", funded by the National Institute of Diabetes, Digestive and Kidney Diseases, U24 DK058739.

Microarrays for an Integrative Genomics

1 Introduction

The functional genomics "meltdown" is coming. At least that is what we fear is likely to occur with the confluence of the high expectations engendered by the Human Genome Project and the prevalence of highly uneven scholarship in the investigations made possible by comprehensive genomic measurement technologies, such as microarray-based expression profiling. Because the availability of these technologies has preceded the development of a substantive canon of appropriate analytic techniques, safe experimental design, and cautionary tales, the quality of these investigations and the manner in which they have been reported often results in the dissemination of highly preliminary and flimsy findings. The absence of widespread use of computational and biological validation procedures associated with these studies has led to many reports that will likely not be substantiated by follow-up studies. With the rapid decrease in the cost of genomic techniques, follow-up studies are likely to arise within the next few years. At that time, several previously well-publicized findings in basic biology, clinical diagnosis, clinical prognosis, and pharmacogenomic targeting will be found deficient. The inevitable reaction to these unfortunate developments will be attenuated if the discipline of functional genomics has matured enough by then to achieve the rigor required.

In this book, we have attempted to address some of the challenges required to begin this maturational process in the intersection between microarray expression technology, bioinformatics, and biomedical science. Despite the likelihood of an eventual disappointment and subsequent retrenchment of the ambitions of microarray applications, we are confident that the development of systematic approaches to this discipline will ultimately deliver on the exuberant predictions and promises made over the last 5 years. We hope that our experience (including our own generous share of mistakes), as communicated here, will help provide the reader with a framework to participate productively in attaining this goal.

1.1 The Future Is So Bright...

Let us be clear lest the above suggest that we are pessimistic about this field. There are very few disciplines within biomedical research with as much promise and excitment as functional genomics. Consider the example of the analysis of large B-cell lymphoma, a deadly malignancy of the lymphatic system, conducted by Alizadeh *et al.* [3]. In (a necessarily abbreviated) summary of their study, the gene expression of the lymphatic tissues of a cohort of patients with the diagnosis of large B-cell lymphoma was measured using DNA microarray technology. That

is, thousands of genes expressed in these tissues were simultaneously measured in the respective lymphatic tissue sample of each patient.

When a clustering analysis was performed to see which patients ressembled one another the most, based on their gene expression pattern, two distinct clusters of patients were found. When the investigators examined the patients' histories it became apparent that the two clusters corresponded to two populations of patients with dramatically different mortality rates (illustrated by the survival curves in figure 4.13, section 4.11).

The implications of these two significantly distinct mortality rates are profound. First, these investigators have discovered with genomic technologies and bioinformatics analyses a new subcategory of large B-cell lymphoma, a new *diagnosis* with clinical significance. Second, they have generated a tool that provides (pending confirmation in other studies) a new *prognosis*; patients can be given much more precise estimates of their longevity. Third, it provides a new *therapeutic opportunity*; patients with an expression pattern predicting a poor response to standard therapy may be treated with different (*e.g.*, much more aggressive) chemotherapy. Fourth, it presents a new *biomedical research* opportunity; what is it about these two subpopulations that makes them so different in outcome and how can that be related to the differences in gene expression?

It is rare that a set of measurement and analytic techniques can so revolutionize biomedical research and clinical practice. It is precisely because the excitement and expectations surrounding this field are so high that we are compelled to inject a note of skepticism about the measurement techniques and the analytic methods. Without such skepticism, the scientific method cannot function and the field will not advance. Nevertheless, even as we have discovered for ourselves significant drawbacks in genomic methodologies, which we address in this book, we remain convinced that these problems only represent the transient growing pains of a new field of biomedical investigation.

Who is this book intended for? Answering this question has served as our constant compass throughout the book's writing. There are three audiences that we have had in mind.

1. *Experienced biologists with limited experience using expression microarrays, or who are concerned that their current approaches to this field are problematic.* For them, this book provides a systematic approach to using microarrays as a tool to investigate biology. Our particular goal for this audience is to realize the pitfalls and *caveats* relating to expression microarray technologies and their analysis. Also, we intend this book to provide these biologists with a firm foundation on which they

can engage in collaborations with colleagues formally trained in bioinformatics and biostatistics. We deliberately limited the amount of mathematical treatment of this material. Where it is present, the reader can skip it without significantly reducing the comprehension of the subsequent text.

2. *Experienced informaticians with limited experience analyzing microarray data.* We believe the field of functional genomics to be one of the most fruitful and rewarding for a computer scientist or other quantitatively trained scientist to enter. It provides a source of challenges, problems, and data sets that will stimulate basic methodological development while furthering important goals in the enterprise of biological discovery and the state of the art of clinical care. For this reason, we emphasize an approach that is driven by the questions of interest to these latter goals. As a result, this book will not present the fundamental computer science underlying various techniques (*e.g.*, proofs of the soundness of various machine-learning algorithms) but will instead illustrate their application to problems that are challenging investigators in functional genomics. This is not to say that the approach we have taken is not rigorous; it just eschews details on the methodologies that are available elsewhere and that we have copiously referenced.

   We are convinced that the most productive research projects will be those that involve intimate collaborations with biologists. This is in contrast to the remote and *post hoc* analysis of the outcome of experiments that the bioinformatician has little to do with, but which is frequently the norm in this discipline. We intend and hope this book can serve as the basis for collaborations in which the informatician understands the goals of biology in this scientific enterprise and in which she or he can contribute to the experimental design and analysis as a first-class member of the research team.

3. *Students entering the field of Bioinformatics.* In our own classes (*e.g.*, Medical Computing 6.872 taught at MIT), we have been gratified to note the emergence of a new generation of students who are both facile in the use of computers as experimental tools and who have a broad understanding of the biological sciences. Although this text can be used as the basis of a course, it is also designed for independent study. For those students, this text is intended to serve as a rapid introduction to the fields of microarray-driven studies of functional genomics so that they can become productive researchers even while they pursue their studies. Indeed, we have had several successful collaborations with undergraduates who have used this material as a launching pad for graduate research or careers in industry.

1.2 Functional Genomics

Now that the human genome has been sequenced (or, more accurately, now that a handful of human genomes have been sequenced), we are said to be in a postgenomic era [200]. We find this term confusing ("genome, we hardly knew ye") because in our view, now that we know at least the draft outline for the genome of multiple organisms, we can begin for the first time to systematically deconstruct how the genetically programmed behavior of an organism's physiology is related to the constituent genes that make its individual version of its species' genome. In this deconstruction, several kinds of biological information are available: DNA sequence, physical maps, gene maps, gene polymorphisms, protein structure, gene expression, protein interaction effects,[1] and a vast literature in MEDLINE [129].

As we use it in this book, functional genomics refers to the overall enterprise of the deconstruction of the genome to assign biological function to genes, groups of genes, and particular gene interactions. These functions may be directly or indirectly the result of a gene's transcription. Much of functional genomics has been and will continue to be the kind of hypothesis-driven biological research pursued for the past decades.

In this book, we address a computationally intensive branch of functional genomics that has emerged as a result of the practical implementation of technologies to assess gene expression thousands of genes at a time. The ability to comperhensively measure expression affords an opportunity to reduce our dependence on *a priori* knowledge (or biases) and allow the organism to point us in potentially fruitful directions of investigation. That is, we describe a hypothesis-generating effort which, if carefully crafted, can then lead to a highly productive set of investigations using more conventional hypothesis-driven research. In this we have been inspired by the work of Arkin *et al.* [12], as have others. Starting from the raw time-series measurements of the substrates of glycolysis (see figure 1.1), Arkin *et al.* were able to computationally reconstruct the glycolytic pathway (figure 1.2). This reverse engineering of the metabolic pathway without prior knowledge is in contrast to the decades of exacting hypothesis-driven elucidation of this pathway by biochemists. It turns out that, as we will discuss later, in experimental design (section 2.1.3), this metaphor is flawed but it remains an icon for one of our major goals: to use bioinformatics applied to functional genomics data to create and re-create the kind of biological pathway charts that are common in most basic biology laboratories.

[1] For those of you who are unfamiliar with what these are, we touch on defining these terms in section 1.5 and in the glossary.

Furthermore, we note that genomic data can be fruitfully exploited even without the assignment of function: a prognostic test for rejection of transplanted kidneys based on the expression level of three genes is useful even if the function of those three genes is poorly known or not known at all.

1.2.1 Informatics and advances in enabling technology

Gene expression detection microarrays are notable not because they can uniquely measure gene expression. There certainly have been many technologies that have allowed for the quantitative or semiquantitative measurement of expression for well over two decades. What distinguishes gene expression detection microarrays is that they are able to measure tens of thousands of genes at a time and it is this *quantitative* change in the scale of gene measurement that has led to a *qualitative* change in our ability to understand regulatory processes occurring at the cellular level. Figure 1.3 provides perhaps the best motivation for the application of information sciences to the functional genomics enterprise. Since DNA sequencing was invented 25 years ago, the number of gene sequences deposited in international repositories, such as GenBank, has grown exponentially, culminating with the entire human genome being sequenced in 2001. Distinguished from this, the amount of *knowledge* about these genes (as measured by the proxy of the number of papers published in biomedicine) has also been growing exponentially, but at a *much* slower rate. As shown, the number of GenBank entries has fast outstripped the growth of MEDLINE. As such, it serves as a proxy for the large gap that has just opened up between our knowledge of the functioning of the genome and raw genomic data. And GenBank entries are just a fraction of the various kinds of data, listed above, generated as part of our investigation of the genome. This volume of data must somehow be sifted and linked to the biological phenomena of interest. Doing so exhaustively, reliably, and reproducibly is a plausible strategy only with the application of algorithmic implementations on computers. This has led to an unprecedented demand for investigators with the knowledge of successful manipulation and analysis of large data sets. These skills may come from education and training in computational physics, chemical engineering, operations research, or financial modeling, but once they are applied to the domain of functional genomics, they can be collectively described as belonging to the domain of bioinformatics.

One reason why the number of known sequences is growing so much faster is the discovery and use of many automated techniques, such as automated sequencers and shotgun sequencing methods. Until the recent advent of gene expression microarrays, we did not have a similar technique to automate the acquisition of knowledge about these genes' behavior in cellular physiology. The past 5 years

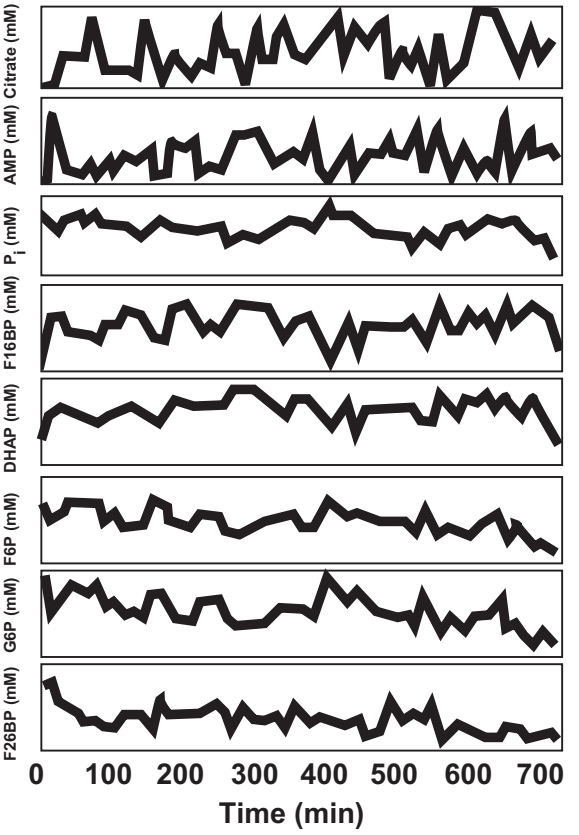

Figure 1.1
Time-series data from anaerobic metabolism. The time courses of measured concentration of the small molecule inputs, adenine monophosphate (AMP), a source of chemical energy to catalyze reactions) and citrate (a substrate), in the experiments, with the responses of the concentrations of phosphate (P_i, an inorganic ion) and of the substrates fructose-1,6-biphosphate (F16BP), dihydroxy acetone phosphate (DHAP), fructose-6-phosphate (F6P), glucose-6-phosphate (G6P), and fructose-2,6-biphosphate (F26BP). (Derived from Arkin *et al.* [12].)

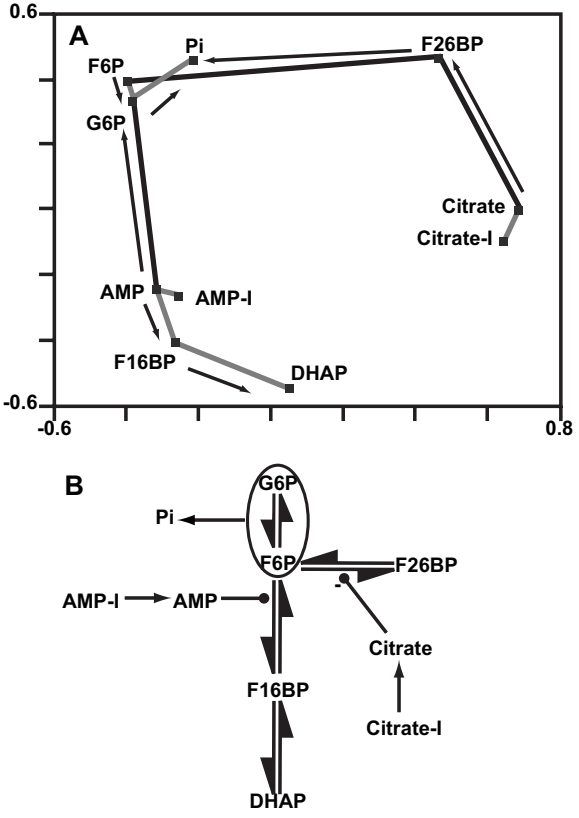

Figure 1.2
Glycolytic pathway reconstructed *ab initio* from time-series data. **A**, The two-dimensional projection of the correlation metric construction (CMC), defined by Arkin *et al.*, for the time series shown in figure 1.1. Each point represents the time series of a given species. The closer two points are, the higher the correlation between the respective time series. Black (gray) lines indicate negative (positive) correlation between the respective species. Arrows indicate temporal ordering among species based on the lagged correlations between the their time series. **B**, The predicted reaction pathway-derived CMC diagram. Its correspondence to the known mechanism of glycolysis is high. (Derived from Arkin *et al.* [12].)

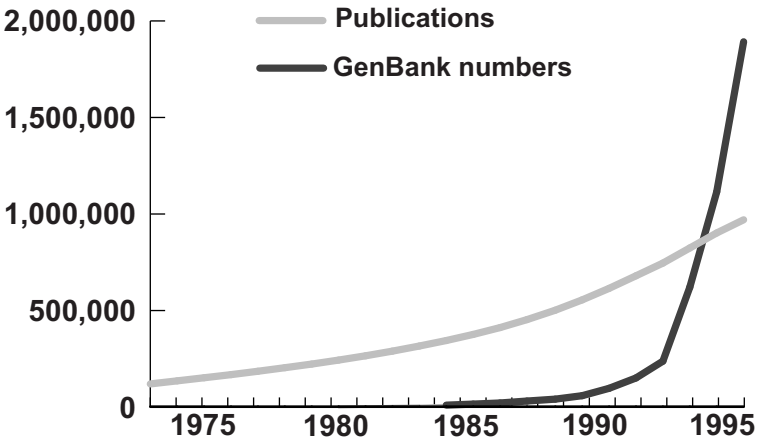

Figure 1.3
Relative growth of MEDLINE and GenBank. The industrialization of genomic data acquisition has created a growing gap between knowledge—of which MEDLINE publications are a proxy—and the information we have gathered about the genome. Information science applied to this information (*i.e.*, bioinformatics) is one of the pillars of an international strategy to overcome this knowledge gap. Cumulative growth of molecular biology and genetics literature (light gray) is compared here with DNA sequences (dark gray). Articles in the G5 (molecular biology and genetics) subset of MEDLINE are plotted alongside DNA sequence records in GenBank over the same time period. (Derived from Ermolaeva *et al.* [65].)

have seen an incredible confluence of disparate technologies, such as robotics, florescence detection, photolithography, and the Human Genome Project,[2], so that today, biologists can use RNA expression microarray detection technologies to obtain near-comprehensive expression data for individual cells, tissues, or organs in various states. With currently available commercial tools, a single experiment using RNA expression detection microarrays can now provide systematic quantitative information on the expression of 60,000 unique RNAs within cells in any given state. Complementary DNA (cDNA) and oligonucleotide microarray technology[3] cannot only be used to determine the abundance of RNA transcripts. By virtue of their broad reach, these measurement platforms permit a large number of exhaustive comparisons: of transcriptional activity across different tissues in the same organism, across neighboring cells of different types in the same tissue, across groups of patients with and without a particular disease or with two different diseases. These platforms can also be used to analyze complex systems, such as traits with multigenic origins or those linked to the environment. They can be used in time series to measure how a particular intervention may start a transcriptional program, *i.e.*, change the expression of large numbers of genes in a reproducible pattern determined by inherent genetic regulatory networks. With sufficient data, they can be even used to provide insight into the underlying mechanisms of these genetic regulatory networks.

Nonetheless, the tools to extract *knowledge* from *data* collected from all of these types of experiments are still in their infancy, and novel tools are still needed to sift through the enormous databases of simultaneous RNA expression to find the true nuggets of related function. The application of techniques of information science, computer science, and biostatistics[4] to the challenge of knowledge acquisition from genomic data is commonly known as *bioinformatics*. This appellation applies to the quantitative and computational analysis of all forms of genomic data, including gene sequence, protein interactions, protein folding, and any observable or measurable phenomenon of interest to the biomedical researcher. The breadth of this commonly used definition of bioinformatics risks relegating it to the dustbin of labels too general to be useful of which artificial intelligence, knowledge management, and systems analysis are only among the more recent. In this book our intent is to be

[2] Using the common shorthand title for the international public and private effort to determine the sequence of bases in the human genome.

[3] The characteristics of the various microarray technologies are addressed in chapter 3.

[4] The label applied often seems to be determined more by the training background of the labeler rather than any fundamental characteristic of the analytic technique.

sufficiently specific about the bioinformatics techniques employed that the matter of a sufficiently broad and yet specific definition of bioinformatics is moot.

Over the past 6 years, several approaches have been developed to analyze microarray-generated RNA expression data sets. The central hypothesis (or hope) of these methods is that, with improved techniques in bioinformatics, one can analyze larger data sets of measurements from RNA expression detection microarrays to discover the "true" biological functional pathways in gene regulation, and develop more definitive, sensitive, and specific diagnostic and prognostic characteristics of disease. However, this is only one of many important areas of bioinformatics addressed in this book. Particularly because we are still in the immediate aftermath of the Human Genome Project, many of the basic naming and data management practices of functional genomics remain in flux and are active areas of bioinformatics development. Although this activity is quite distinct from the analytic efforts touched on above, it currently consumes perhaps the largest proportion of the bioinformatics community because its resolution is urgent and a *sine qua non* for the success of any of the analytic efforts. After all, if we cannot reliably name the same gene in identical fashion across experiments, if we cannot reliably retrieve expression data from all the microarray experiments of interest, if we cannot readily access the meaning and function of genes determined by thousands of researchers, then the whole enterprise of functional genomics will be crippled if not intractable.

There is a related discipline to bioinformatics—*clinical informatics*—which refers to the application of information science to various aspects of clinical care. Although clinical informatics is not addressed in this book in detail, in chapter 6 we describe many of the problems that have dogged clinical informaticians and that will confront bioinformaticians as they attempt to bring their basic science findings to clinical relevance.

1.2.2 Why do we need new techniques?

A first look at a typical genomic study might cause a quantitatively trained scientist or even a biologically trained scientist to ask the following, quite legitimate question: Why is this field not amenable to standard biostatistical techniques? After all, we are trying to understand the relationship between multiple variables and the mechanisms that the relationships reveal. And there has been a long history of the development of biostatistical techniques to analyze large studies with large numbers of cases with many variables to elucidate precisely this kind of question. Specifically, these studies ask questions such as: What risk factors are associated with heart disease? Does smoking cause disease? What is the difference in survival between a group treated with one chemotherapeutic drug versus another? On

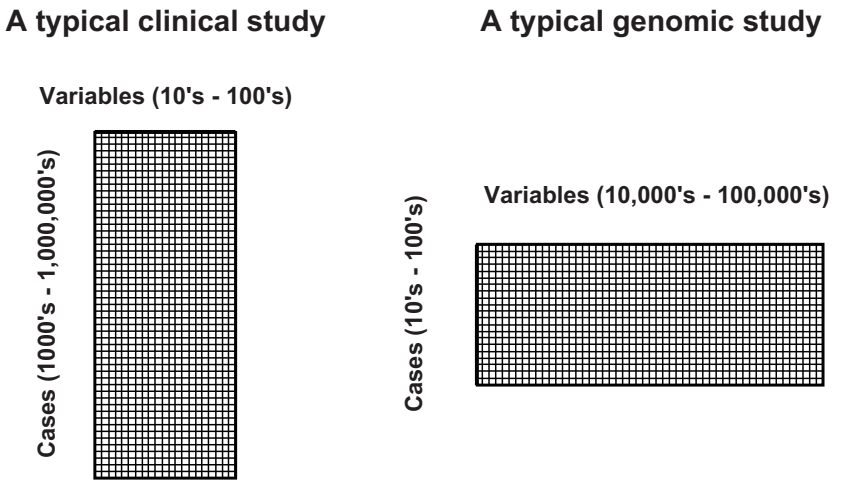

Figure 1.4
A major difference between classic clinical studies and microarray analyses. The high dimensionality of genomic data in contrast to the relatively small number of samples typically obtained results in a highly underdetermined system.

the surface these questions seem similar to many of those posed regarding genetic risk factors for acute and chronic disease. Yet a review of the bioinformatics and functional genomics literature over the past 3 years reveals that most of the analyses have been performed using techniques borrowed from the computational sciences and machine-learning communities in particular. Why is this? There are several reasons, including academic parochialism, but perhaps the most substantive one is the essentially underdetermined nature of genomic data sets as described below. If we examine figure 1.4, we see sketched out the fundamental difference between a typical clinical study and a typical genomic study. A high-quality clinical study will involve thousands to tens of thousands of cases, such as in the Nurses' Health Study [18] or the Framingham Heart Study [55] over which tens or even hundreds of variables are measured. In contrast, in a typical genomic study, there are only tens or, exceptionally, hundreds of cases, but thousands of measured variables.

Initially, the low number of cases in functional genomic investigations may have been due to the high cost of the microarrays (on the order of several thousand dollars per microarray) but increasingly the scarcity of cases in a typical functional genomics study will relate to the scarcity of appropriate biological samples. As

these experiments involve the measurement of gene expression, a particular tissue has to be obtained under the right conditions. This is in distinction to genomic DNA samples where most blood samples will suffice. Especially in human populations, suitable tissue samples are relatively rare.[5] Yet even though there are only tens of cases, each case involves the measurements of tens of thousands of variables corresponding to the expression of tens of thousands of genes measurable with microarray technology. The result of the large number of variables compared to the number of cases is that we have highly *underdetermined* systems. That is, we are making measurements of very high dimensionality (on the order of tens of thousands) but we are only providing a small number of cases to explore this high-dimensional space. Another way to say this is that there are many, many ways in which the variables being measured could be interrelated mechanistically, based on the relatively small number of observations. Due to this high dimensionality and the underdetermined nature of these systems, standard biostatistical techniques do not hold up well because many of the assumptions that underlie these conventional biostatistical techniques do not hold.[6] We often provide the following analogy. To solve a linear equation of one variable (*e.g.*, $4x = 5$) we only need one equation to find the value of the variable. To solve a linear equation of two variables (*e.g.*, $y = 4x + b$), two equations are required. If we have tens of thousands of variables, but only hundreds of equations, then there will be thousands of potentially valid solutions. This is the essence of what constitutes an underdetermined system. In this context, we must use techniques that can maximally inform us of the relationships between variables of interest (and find out which ones are of interest) despite the underdetermined nature of the data sets. High-dimensionality data sets have been well-known to the "machine-learning" community of computer scientists in applications such as automated recognition of human faces, so it is not surprising that many of the techniques developed by that community have found their way into the functional genomics enterprise.

[5] See chapter 2 for a discussion of which tissues are appropriate for particular experiments.

[6] Although this holds true for most of the biostatistical techniques biologists will have learned in graduate school, in fairness there has been quite a lot of research by statisticians on the analysis of underdetermined systems of high dimensionality. Their work has just not found its way into mainstream biomedical study until recently.

1.3 Missing the Forest for the Dendrograms, or One Aspect of Integrative Genomics

In the first 2 years of significant publications regarding the large scale application of microarray technologies, numerous special purpose or adapted machine-learning algorithms were described in the literature. Self-organizing maps [175], dendrograms [27, 63, 76], K-means clusters [101], support vector machines [33, 72], neural networks [107], and several other methodologies (borrowed largely from the machine-learning community of computer science) have been employed. Most of these have worked reasonably well for the purposes described in the papers. It is one of the central contentions of this book, however, that the choice of a particular clustering or classification methodology is secondary to proper experimental design and full knowledge of the properties and limitations of massively parallel expression analysis in general and those of the specific microarray technologies employed. This contention does not constitute an overly fastidious approach to functional genomics but the insight from (often expensive) experience gleaned through our own mistakes and those of our colleagues and the reported investigations. The bioinformatics technique centric approach has been evident in our own collaborations. Often, at the outset of a new investigation, our collaborators from both the computational community and the biological community immediately wish to address the questions of which are the appropriate clustering techniques and which one is "the best." While we certainly recognize that the choice of clustering or classification technique is important (and we devote chapter 4 to this matter), we are firmly convinced that it is only one part of a well-designed pipeline that defines a successful exploration of biology and medicine using microarray technology. This pipeline is diagrammed in figure 1.5 on page 17. We refer to this functional genomics pipeline throughout the book. We discuss at length the practicalities of assembling this pipeline in subsequent chapters, but a few characteristics of this pipeline bear mentioning here:

- *Selection of the right tissue.* Experiments in functional genomics require selection of the functionally relevant tissue or cell type. In certain experiments, like those using blood and solid cancers, the functionally relevant tissue is clear. In other analyses, the functionally relevant tissue is not so easily ascertainable or acquirable. For example, the clinical phenotype seen in type 2 diabetes mellitus, or insulin resistance, involves the coordinated physiological dysfunction of several organs and cell types, including liver, muscle, and fat cells. Schizophrenia involves a higher-order brain dysfunction, but brain cells are not easily accessible in humans. For some common

diseases like hypertension, it is not clear what the functionally relevant tissue is. A successful pipeline involves collaboration with a source of tissue, such as a surgical team, a laboratory with biologically interesting animals, or a laboratory with cell lines of interest.

- *Right conditions.* Even if the appropriate tissue is selected from the organism of interest, the conditions under which the tissue is obtained (*e.g.*, number of hours post mortem) can determine whether or not the investigation is successful. An insulin-sensitive tissue such as skeletal muscle will have a different characteristic metabolic and expression profile depending on the glucose and insulin concentrations prior to the extraction of RNA. The time of day will influence the expression of genes in all tissues which have endogenous circadian rhythms or that have processes that can be entrained by physiological clocks. Awareness of these issues and cooperation from a surgeon, pathologist, or technician responsible for obtaining the tissue is therefore an essential component to the success of the functional genomic pipeline.

- *Extracting RNA, hybridizing to microarray, and scanning.* Each of these steps in this "wet" component of a functional genomics pipeline is susceptible to operator error and is a potential source of poor or noisy measurements. The RNA extracted may be of poor quality, the hybridization conditions may vary (*e.g.*, the room temperature), and the settings of the scanner that produces the digital image of the microarray may vary from one scan to another. Industrialization and standardization of this component has been the focus of the more successful and high-quality functional genomics efforts using expression microarrays.

- *Functional clustering.* This "dry" component of the pipeline is often thought to be what bioinformatics is about. And in fact, it may be at this stage that the algorithmic analysis of an expression profiling study to detect biologically or clinically meaningful patterns or associations is the only time a bioinformatician is involved. We will argue throughout this book that a successful functional genomics pipeline involves the bioinformatician at every step.

- *Computational validation.* As will be elaborated in this book, there are many reasons to perform bioinformatics analyses on functional genomics data sets, and many methods can be used. One unique problem with these types of data sets is that they are "short and wide," meaning that many characteristics are measured on relatively few samples. For example, current microarrays offer the quantitation of up to 60,000 expressed sequence tags (ESTs) in any given sample, but current costs may limit a single experiment to 10 to 100 samples. Because of this problem, these

data sets are essentially underdetermined, as described on page 10, meaning that there are many correct ways to mathematically describe the clusters and genetic regulatory networks contained within them. Thus, some computational validation is required immediately after the bioinformatics analysis so that computationally sound but biologically spurious or improbable hypotheses are screened out.

The principal motivation for the screening out of spurious or improbable hypotheses is the efforts that follow. Each hypothesis generated that passes this step may need to be validated in a biological laboratory. Some biological laboratories may wish (and may have the resources) to pursue many hypotheses and can tolerate the eventual refutation of large numbers of false-positive hypotheses. Other biological laboratories may only be able to validate a few. Thus, a proper bioinformatics analysis includes a computational validation. An ideal computational validation does not merely provide a yes or no answer as to the potential validity of a hypothesis, but instead provides a continuum of validation, or a receiver-operating characteristic curve. With such a curve, the biologist can select the desired point of sensitivity or specificity and true and false negatives and positives (see sections 2.1.4 and 4.12.3).

- *Biological validation.* Most biological questions will not be answered using microarrays. Instead, the most likely outcome from a functional genomics analysis is the next biological question to ask. As hypotheses are generated from bioinformatics analyses, biological validation is crucial to verify these hypotheses. This verification may include, for instance, making sure a particular set of genes is truly expressed at the proper time and place as hypothesized, using conventional biological techniques such as Northern blotting and *in situ* hybridization.

- *The multidisciplinary team.* In most settings, all of these steps, from acquisition of source material, to microarray construction, to bioinformatics analysis, to biological verification, cannot be performed by a single group or laboratory. A successful functional genomics pipeline brings together resources from many disciplines and of varied backgrounds. Two anecdotes serve to illustrate the value of this multidisciplinary approach.

  We were in the process of analyzing a large number of microarray expression data obtained from skeletal muscle for some colleagues interested in muscular dystrophy—a class of genetic diseases of muscle. They were gratified when our clustering analyses found interactions between transcriptional factors and contractile proteins that they had discovered just months before using conventional molecular-biological techniques as well as several new but plausible interactions. However, be-

cause the clustering analyses were exhaustive, they also identified several hormonal interactions that were not of primary interest to these neuromuscular specialists. Using annotation tools linking the microarray data to several national databases, it became quickly apparent that these hormonal interactions were thought to be exclusive to adipocytes (cells constituting the principal component of fatty tissue) but we had just found suggestive evidence to the contrary. The multidisciplinary nature of our effort allowed the formulation of well-posed questions directly related to the interests of the biological investigators and yet kept us open to important hypotheses generated from the data.

We are participating in a study of the functional genomics of the developing brain using mouse models. We had computed, using approximately two dozen microarray data sets produced by our collaborators—developmental biology researchers—a list of approximately 100 genes that appeared to be involved in the development of a specific region of the brain. Our collaborators were in the process of selecting a subset of these for biological validation but we were worried about the outcome of the validation because the data had been derived from entire portions of the brain. whereas the process the developmental biologists were interested in only occurred in a minute component of the brain. It seemed probable that many of the 100 genes were not specific to the processes we were studying. Given only the expression data from the microarrays, none of the bioinformatics techniques were able to further refine or hone the list of 100. Fortunately, the developmental biologists provided us with the following insight. They knew of one gene g that was expressed in the tiny area of the brain that they were studying and they had determined empirically that it was expressed in no other part of the brain. They suggested that we find all those genes in the list of 100 that behaved very closely to g. We found 10 such genes and our collaborators went on to successfully validate 8 of them using the techniques of conventional hypothesis-driven molecular biology. If we had not drawn on the multidisciplinary capabilities of our team for that small but crucial biological insight, then we would have been stuck with a large list of nonspecific genes of little relevance to the questions originally posed by the developmental biologists.

When the initial design of the multidisciplinary functional genomics pipeline is given short shrift and the fundamental limitations of expression microarray technologies misunderstood, the enterprise of functional genomics appears to approximate the "fishing expedition" that has been the oft-stated concern of traditional biologists regarding this nascent discipline. Consequently, even if a particular investigator participates in only a fraction of the pipeline, understanding the safe

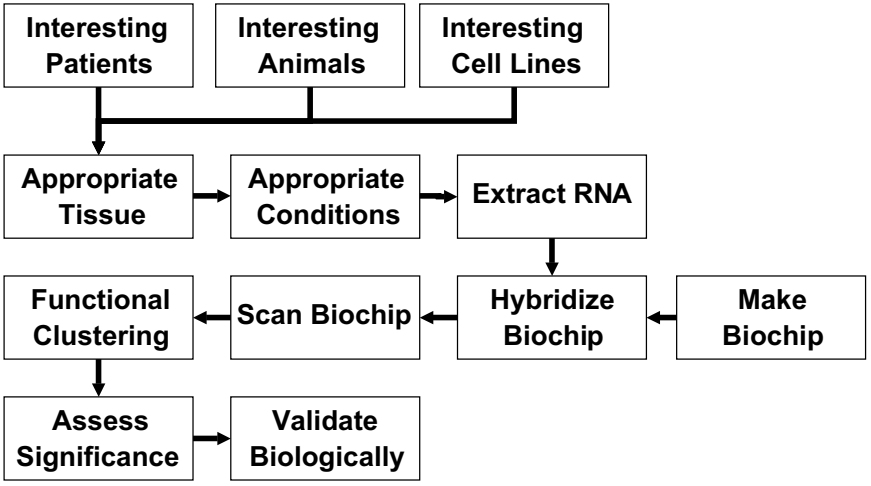

Figure 1.5
An archetypal functional genomics pipeline. Shown is a simplified view of a functional genomics pipeline solely involved in expression microarray experiments. Note the interdigitation of "wet" and "dry" components requiring close multidisciplinary collaboration and some creative consideration of the value of the individual contributions in this pipeline for a particular experiment and publication.

design of an entire functional genomics pipeline can maximize the yield of these experiments, or at the very least produce convincing and reproducible negative results. It is the intent of this book to point the way to investigations that provide such an understanding.

Because of the dramatically different backgrounds (at least at present) of the different contributors to the functional genomics pipeline, its social dynamics may be challenging, as described below. We pay attention to these dynamics because one aspect of an *integrative genomics* is the integration of disciplines and experts (the other side will be described before the end of this chapter).

1.3.1 Sociology of a functional genomics pipeline

Given the requirement for a multiinvestigator, multi-institutional effort, several pragmatic decisions and realities must be confronted. We describe these in the context of academic collaborations but analogies to the commercial world are obvious. The first question in academic collaborations around the functional genomics pipeline often is: Who will get the credit for the work and for the discoveries that ensue from a particular functional genomics investigation? More concretely, who will get first and senior authorships in the publications that report on the discoveries obtained from this functional genomics discovery pipeline? Will it be the surgeons who obtained the tissue, the bioinformaticians who performed the cluster analysis, the biologist who runs the microarray facility, or the clinician who obtained the phenotypic characterization of the patient from whom the tissue was obtained?

Resolving this issue is nontrivial because of some fundamental human and cultural considerations of the various types of investigators. Within each discipline, an investigator will tend to see those outside his or her discipline as performing more of a utility function rather than making a significant intellectual and creative contribution to the research process. For example, a molecular biologist might view the bioinformatician as providing cookbook analytic algorithms to winnow out the relevant findings from a biological system that they have spent time, energy, and creativity in developing. Conversely, the bioinformatician might view the biologist as a laborer plodding along in the murky swamps of biological experimentation on which the bioinformatician can then shed light by virtue of his or her insights into the general principles of automated inference, clustering, and classification.

For those of us who have labored on both sides of this cultural divide, it is quite clear that there are plodders and creative geniuses in both fields. Successful collaborations in this arena require thoughtful recognition of relevant contributions prior to initiating a long-term collaboration around this strategy. In our own collaborations, we have adopted the following heuristics: If the bioinformatics techniques

used are not novel and the interpretation is rote, then the biologists and clinicians usually assume first and senior authorships. If the experimental design is innovative and informed by the nature of the bioinformatics analyses, and the computational techniques have to be developed or customized, then the bioinformaticians usually take first and senior authorships. Often, the major contributions are mixed, as when the bioinformatics analyses are novel and the biological validation steps are creative and arduous, in which case the authorships are split accordingly. In the end, however trite as it sounds, it is true that nothing substitutes for the collegiality and good will that arise from mutual respect for different skills and contributions.

1.4 Functional Genomics, Not Genetics

We have noticed, among some of our biological collaborators, a tendency to view the massively parallel methods of functional genomics as a highly efficient large-scale application of methods that they have already applied. For instance, gene expression profiling and polymerase chain reaction (PCR) are all methods that have been used by molecular biologists for decades. What we hope the reader will obtain from this book is an appreciation of how the near-comprehensive (and soon to be truly comprehensive) nature of the functional genomics approach as permitted by expression microarrays changes qualitatively and fundamentally the nature of biological investigation. Before our potential readers with biological backgrounds become offended by or disgruntled with this assertion, let us assure them that we present an equivalent critique for the purely computationally oriented bioinformatists and genomicists in the following section. Functional genomics is not, as some have portrayed it, a hypothesis-free fishing expedition, nor is it, even more charitably, only a hypothesis-generating enterprise requiring subsequent biological validation. It is fundamentally different in that it permits the posing of large questions that are grounded in an essential biological understanding of a particular domain. Unlike the questions posed in "traditional" genetics or molecular biology, these questions have less stringent requirements for prior supposition or claims of the role of a particular gene or metabolite in a biological process. An example of some of the broad questions that can be asked are:

- Which of all the known genes have regulatory mechanisms that appear to be similar to those regulated by the sonic hedgehog transcription factor in the cerebellum?
- Given the effect of 5000 drugs on various cancer cell lines, which gene singly is the most predictive of the responsiveness of the cell line to any chemotherapeutic agent?

- Given a known clinical distinction, such as that between acute lymphocytic leukemia and acute myelogenous leukemia, what is the minimal set of genes that can most reliably distinguish these two diseases?
- Is there a group of genes that can serve to distinguish the outcomes of patients with large B-cell lymphoma that are otherwise clinically indistinguishable on presentation?
- What distinguishes the signaling pathways of two of the substrates of the insulin receptor?

These questions are important biologically and clinically, and yet they can only be posed reasonably if they involve a comprehensive view of genomic regulation and involve the use of computational methods that can efficiently sift through the vast quantities of genomic data generated by the experiments required to answer these questions. Another way to consider functional genomics is to view it as serving as a filtered funnel through which these broad questions can be strained. The residue that remains is high-yield, detailed, and contains particular biological questions that are answerable by more traditional genetic or molecular biology techniques. This is illustrated in figure 1.6 below. The utility of this metaphor is as follows. The universe of possible participants in any given biological regulatory mechanism is finite but very large. Even with the most comprehensive in-depth expertise, a biologist may find be surprised about insights obtained through data mining of expression patterns, human genome sequences, and often from data obtained from other species. Without a genomic approach to guide her experiments, this biologist may expend several months in false leads or alternatively miss an important component to the system under study. Similarly, if the biologist is looking at a set of genes that are thought to be predictive of a given clinical condition, such as transplant rejection or cardiac disease, without the comprehensive view brought by functional genomics, elements of the diagnostic or prognostic procedure, such as the concentration of a gene transcript or a protein, may be omitted with a concomitant decrease in the sensitivity and specificity of the prognosis or diagnosis.

1.4.1 *In silico* analysis will never substitute for *in vitro* and *in vivo*

There is little doubt that one of the tremendous accomplishments of the Human Genome Project is that it has enabled a rigorous computational approach to identi-

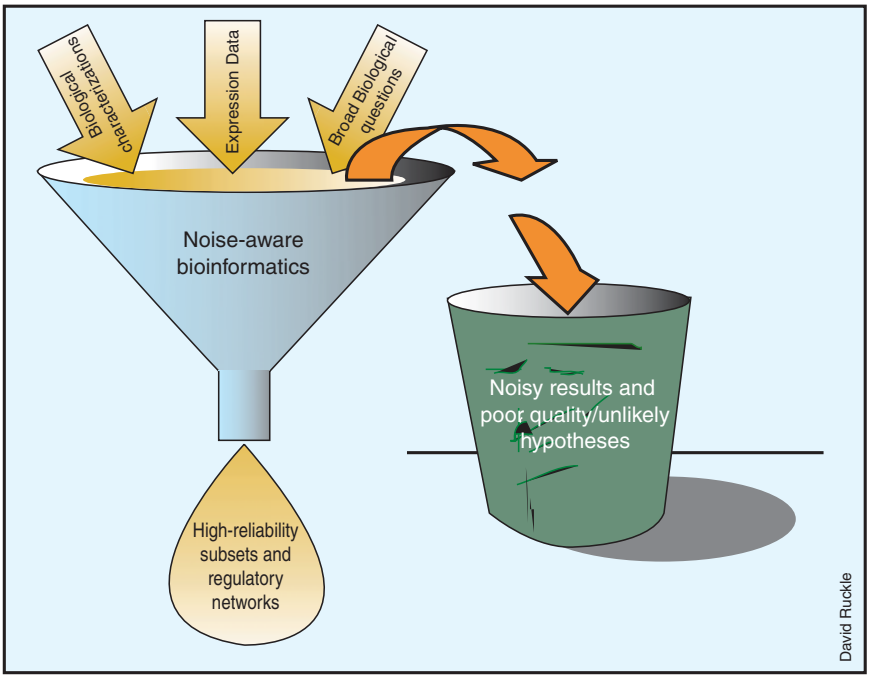

Figure 1.6
The functional genomics investigation as a funnel for traditional biological investigations. Broad questions and comprehensive data are the mix in which bioinformatics techniques are filtered to separate high-yield hypotheses or candidate genes from spurious findings and poor-quality hypotheses.

fying many questions of interest to the biological and clinical community at large.[7] However, the danger of a computational triumphalism is that its makes several dubious assumptions. The first is genetic reductionism: At an abstract level most bioinformaticians understand that a particular physiology or pathophysiology is the product of the genetic program created by the genome of an organism and its interaction with the environment throughout its development and senescence. In practice, however, a computationally oriented investigator often assumes that all regulation can be inferred from DNA sequence, based solely on the syntax of its sequence elements. That is, it is assumed that it is predictable whether a nucleotide change in a sequence will result in a different physiology. We refer to this as "sequence-level reductionism."

The second dubious assumption is the computability of complex biochemical phenomena. One of the most venerable branches of bioinformatics involves modeling molecular interactions such as the thermodynamics of protein folding, and protein-protein and protein-nucleic acid interactions. As yet, all the combined efforts and expertise of bioinformaticians have been unable to provide a thermodynamically sound folding pattern of a protein in the heterogeneous solvent environment of a cell for even as long as one microsecond. Furthermore, studies by computer scientists over the last 10 years [23, 114] suggest that the protein-folding problem is "NP hard." That is, the computational challenge belongs to a class of problems that are believed to be computationally intractable. Therefore it seems overly ambitious to imagine that within the next decade we will be able to generate robust predictive models that are able to accurately predict the interactions of thousands or millions of heterogeneous molecules and the ways in which they modulate the transcription of RNA and the translation of messenger RNA (mRNA) into protein and the subsequent functions of these proteins. We refer to this ambition as "interactional reductionism."

This is not to say that models have no useful role in molecular biology or bioinformatics. On the contrary, they are extremely useful to embody what we currently believe we know about biological systems. Where the predictive capabilities of these systems break down points to where we should guide further research. Also, the educational value of such models cannot be underestimated.

The final questionable assumption is the closed-world hypothesis. Both sequence level reductionism and interactional reductionism are predicated upon the availability of a reliable and predictive and complete mechanistic model. That is, if a

[7] Although even the most basic of the original conclusions, stated in the publications heralding the completion of the draft of the human genome, the order of magnitude of the number of genes in the human genome, now seems to be again in contention.

fertilized ovum can follow the genetic program to create a full human being after 9 months, then surely a computer program should be able to follow the same genetic code to deterministically infer all the physiological events that are determined by the genetic code. Indeed, there have been several efforts, such as the E-cell effort of Tomohita *et al.* [179], which aimed to provide robust models of cellular function based on the known regulatory behavior of cellular systems. Although such models have important utility, our knowledge of all the pertinent parameters for these models appears grossly incomplete today. These parameters are required to describe intracellular processes, intercellular processes, and the unimaginably large repertory of possible environmental interactions with both sets of processes. This incompleteness, and the lack of knowledge of where the boundaries are between the complete and the incomplete, imply that these models will have behaviors that may diverge substantially and unpredictably from those that actually occur.

These caricatured positions of the traditional molecular biologists and the computational biologists are, of course, overdrawn. When prompted, most of these investigators will articulate the fullness of the complexities of the analytic tasks of functional genomics. In the conduct of their research or even in the discussions within their publications, these same investigators will nonetheless often retreat to the simplifications and assumptions described above. This may be because they take for granted that their colleagues and readers understand these simplifications, but such unstated assumptions can often misdirect novices in this discipline.

What we argue for, and hope that this book communicates, is the necessity for a rapid generate-and-test paradigm that cross cuts repeatedly across the disciplines of genetics, computational biology, and molecular biology. Operationally, this means that bioinformatics tools can be used to guide the investigations of an experimental biologist investigating a particular biological system or disease process. But for even the smallest assumption, rather than relying on the statistical association or predicted behavior of a system, empirical evidence has to be developed to support these. It is only in this incremental accretion of evidence that the discipline of functional genomics can become a science.

Why microarrays? Why focus on microarrays? After all, there are many other ways to impute function to genes. Using only genomic DNA easily obtainable from a peripheral blood sample or a buccal smear, genetic epidemiologists can conduct association studies using microsatellite markers or polymorphisms [43] to associate prognoses, diagnoses, and even biological function with a particular gene [67, 126]. And then there are the more conventional genetic techniques of transgenic and misexpression whole-organism models of the function of various genes. Even more

recently, the feasibility studies for proteomic assays suggest that in the future we will directly be able to assess changes in protein concentration at the cellular level.

In contrast to linkage and association studies, microarray studies are designed in principle to measure *directly* the activity of the genes involved in a particular mechanism or system rather than their association with a particular biological or clinical feature. An association-linkage-genetic epidemiology study relies upon a long indirect probabilistic causal chain: that a change in DNA sequence results in a change in gene regulation or protein structure, resulting in a change in cellular physiology measurable as a change in a whole-organism profile (*e.g.*, the human phenotype). For some changes in genomic sequence, particularly in the instances of multigenic regulation, the effects may be so small that any conceivable population study may not be able to detect them.[8] Also, the cost of screening the genome of sufficiently large populations to achieve adequate statistical power has prohibited all but the most focused association studies (although this is likely to change). Unlike the current state of art and engineering of large-scale proteomic assay systems, gene microarrays are currently affordable and within many applications have acceptable reproducibility and accuracy.

Another aspect of an integrative genomics Notwithstanding these apparent advantages of expression microarray studies, as we discuss in sections 1.5.3 and 1.5.2, there are several kinds of information that we are missing by not including such measurements. Our decision to restrict the scope of this book to the exploration of functional genomics and genomic medicine from the perspective of microarray technology is then largely a pragmatic one. Expression microarrays are sufficiently well engineered and cost-effective to allow thousands of researchers to productively employ them to drive their investigations. If, in the future, as we expect, massively parallel measurements of individual proteins becomes cost-effective, large-scale, and highly reproducible, then we will certainly expand the analysis to address these methodologies. The same will be true when high-resolution (*i.e.*, every kilobase) genome-wide scans of hundreds of individuals will become economically feasible for most clinical research studies. Current estimates have these technologies available on the genomic and population scale within 5 years. A well-prepared genomic investigator will have prepared the pipeline to take advantage of all these measurement technologies.

[8] That is, there would be insufficient numbers of individuals with the necessary constellation of phenotypes across the entire human population. At the same time we recognize that there are several diseases, such as sickle cell anemia, where the change of one base in the hemoglobin gene results in a severe and unsubtle disease phenotype.

From the computational perspective, the measurement of any analyte, whether it be an inorganic constituent of serum, an RNA transcript, or a protein, are all simply point measurements of variables corresponding to the total state of the cell. Likewise, all clinical measurements (*e.g.*, height, blood pressure) and history (*e.g.*, age of menarche, time of cancer diagnosis) are point measurements corresponding to the total state of the organism (*e.g.*, human). It is only in the important details of the quality and meaning of these measurements that they differ. This is said both with tongue planted firmly in cheek and in all seriousness.

And indeed, it is this other aspect of an *integrative genomics* to include as many modes of data measurement as are available. Each mode reflects another aspect of cellular and organismal physiology each with its own set of specificities and sensitivities with respect to a phenomenon or process of interest. The role of bioinformatics in an integrative genomics is to both provide the glue bringing together all kinds of genomic and phenotypic data and the means to extract knowledge (or at least high-yield hypotheses for subsequent testing) from them in an efficient, large-scale, and timely fashion. It is also the inspiration for the title of this book.

1.5 Basic Biology

This section is meant solely for the quantitatively trained scientist who knows essentially none of the biology developed over the last three decades. For those of you who need a significant refresher in molecular biology we list several good introductory texts and on-line resources (table 1.1). So we start from the beginning. In almost all cells making up a living organism, there is believed to be an identical set of codes describing the genes and their regulation. This code is encoded as one or more strands of the deoxyribonucleic acid molecule: DNA. That is the same in almost every cell in the human body.[9] For instance, a liver cell and a brain cell have the same DNA content and code in their nucleus. What distinguishes these cells from one another is that portion of their DNA that is transcribed and translated into protein, as described below.

The entire complement of DNA molecules of each organism is also known as its *genome*. The overall function of the genome is to drive the generation of molecules, mostly proteins, that will regulate the metabolism of a cell and its response to the environment, as well as provide structural integrity.

[9] There are several exceptions to this rule, such as mature red blood cells, which lack nuclei and therefore the organism's genome, and gametes (spermatozoa or ova), which have half the usual complement of DNA. However, for the purposes of this overview, the above generalization will suffice.

- *Genomes.*
  T. A. Brown & Austen Brown. New York, Wiley-Liss, 1999.

- *Human Molecular Genetics*, 2nd edition.
  Tom Strachan & Andrew Read. New York, Wiley-Liss, 1999.

- *Primer on molecular genetics*
  `http://www.bis.med.jhmi.edu/Dan/DOE/intro.html`.

- *Primer on genomics with a commercial flavor*
  `http://www.biospace.com/articles/genomics.primer.cfm`

- *Introductory biology course at MIT (7.01) hypertext book*
  `http://esg-www.mit.edu:8001/esgbio/701intro.html`

Table 1.1
Molecular biology primers

Structure of DNA Each molecule of DNA may be viewed as a pair of chains of the nucleotides adenine (A), thymine (T), cytosine (C), and guanine (G). Moreover, each chain has a polarity, from $5'$ (head) to $3'$ (tail). The two strands join in opposing polarity ($5'$ binds to $3'$) through the coordinated force of multiple hydrogen bonds at each base-pairing, where A binds to T and C binds to G. [10]

DNA is able to undergo duplication, which occurs through the coordinated action of many molecules, including *DNA polymerases* (synthesizing new DNA), *DNA gyrases* (unwinding the molecule), and *DNA ligases* (concatenating segments together).

Transcription of DNA into RNA In order for the genome to direct or effect changes in the cytoplasm of the cell, a transcriptional program may be activated for the purpose of generating new proteins to populate the cytosol—the heterogenous intracellular soup of the cytoplasm. DNA remains in the nucleus of the cell, while most proteins are needed in the cytoplasm of the cell, where many of the cell's functions are performed. Thus, DNA is copied into a more transient molecule

[10] These pairings are present in all DNA and are the most thermodynamically stable of all possible pairings of nucleotides, which accounts for the high specificity with which complementary strands of nucleotide polymers bind to each other.

called RNA. [11] A *gene* is a single segment of the coding region that is transcribed into RNA. RNA is generated from the DNA template in the nucleus of the cell through a process called *transcription.*[12]

The RNA sequence of base pairs generated in transcription corresponds to that in the DNA molecules using the complementary A-T, C-G, with the principal distinction being that the nucleotide uracil is substituted for the thymine nucleotide. Thus, the RNA alphabet is $ACUG$ instead of the DNA alphabet $ACTG$. Each cell contains around 20 to 30 pg of RNA, which represents 1% of the cell mass. The RNA that codes for proteins is called *messenger RNA*, and the part of the DNA that provides that code is called an *open reading frame* (ORF). When read in the standard $5'$ to $3'$ direction, the portion of DNA before the ORF is considered *upstream*, and the portion following the ORF is considered *downstream.*

The specific determination of which genes to transcribe is determined by *promoter regions*, which are DNA sequences upstream of an ORF. Many proteins have been found containing parts that bind to these specific promoter regions, and thus activate or deactivate transcription of the downstream ORF. These proteins are called *transcription factors.*[13]

A diagram of the genetic information flow, from DNA to RNA to protein, is illustrated in Figure 1.7

Prokaryotic and eukaryotic cell types Although there are many taxonomies one could use, we can essentially divide the world of organisms into two types: *eukaryotes*, whose cells contain compartments or organelles within the cell, such as mitochondria and a nucleus; and *prokaryotes*, whose simpler cells do not have these organelles. Animals and plants are examples of eukaryotes, while bacteria are prokaryotes. Most prokaryotes have a smaller genome, typically contained in a single circular DNA molecule. Additional genetic information may be contained in smaller satellite pieces of DNA, called plasmids.

[11] That portion of the entire DNA molecules that is transcribed into RNA is called the *coding region.*

[12] Transcription involves unwinding a DNA molecule so that the particular gene that is to be transcribed is sufficiently exposed to the transcriptional machinery, notably RNA polymerase.

[13] Not all RNA codes for proteins, however. In fact, only 4% of total RNA is made of coding RNA. Of the noncoding RNA, ribosomal RNA (rRNA) and transfer RNA (tRNA) are used in various components of the protein translational apparatus mentioned below, and are not themselves translated into proteins. Eukaryotes also contain small nuclear RNA (snRNA), which is part of the splicing apparatus (see below); small nucleolar RNA (snoRNA), which is involved in methylation of rRNA; and small cytoplasmic RNA (scRNA), which can play a role in the expression of specific genes.

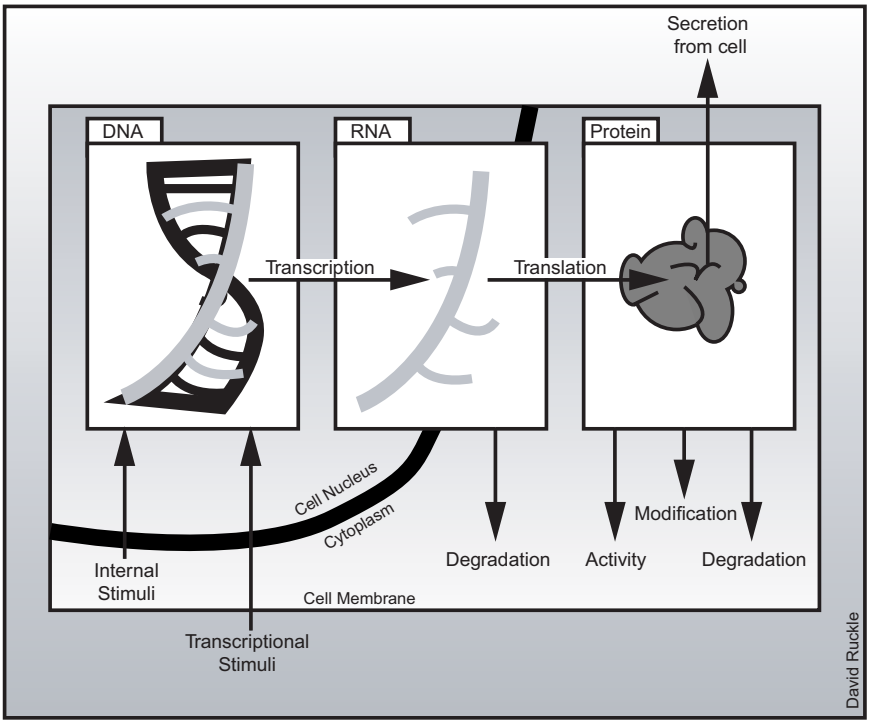

Figure 1.7
Flow of genetic information, from DNA to RNA to protein. This simplified diagram shows how the production of specific proteins is governed by the DNA sequence through the production of RNA. Many stimuli can activate the specific transcription of genes, and proteins can play a wide variety of roles within or outside cells. Note that even in this simplified model, it is obvious that since we are currently able to (nearly) comprehensively measure only gene expression levels, we are missing comprehensive measurements of protein modification, activity, transcriptional stimuli, and many other components of the state of a cell.

The structure and processing of RNA transcripts Eukaryotic genes are not necessarily continuous; instead, most genes contain *exons* (portions that will be placed into the mRNA) and *introns* (interruptions that will be spliced out). Functions have been recently discovered for introns, such as promoter-like control of the transcription process. Introns are not always spliced consistently; if an intron is left in the mRNA, an *alternative splicing product* is created. Various tissue types can flexibly alter their gene products through alternative splicing. [14]

Before coding RNA is ready as mRNA, the pre-mRNA must be processed. [15] In eukaryotes, after the splicing process, the generated mRNA molecule is actively exported through nuclear pore complexes into the cytoplasm. The cytoplasm is where the cellular machinery, in particular the ribosomal complex,[16] acts to generate the protein on the basis of the mRNA code. A protein is built as a polymer or chain of amino acids, and the sequence of amino acids in a protein is determined by the mRNA template. The mRNA provides a degenerate coding in that it uses three nucleotides to code for each of the twenty common naturally occurring amino acids that are joined together to form the polypeptide or protein molecule. With three nucleotides there can be 4^3 possible combinations for a total of 64 combinations. Consequently, several trinucleotide sequences (also known as codons) correspond to a single amino acid. There is no nucleotide between codons, and a few codons represent start (or initiation) and stop (or termination).[17] The process of generating a protein or polypeptide from an mRNA molecule is known as *translation*.

As an example, if an RNA transcript had the nucleotide sequence

GCT TGC AAG GCG

(the spacing and grouping by three, which obviously is not visible in the RNA transcript, was added to make this easier to read), the first three nucleotides would code for this chain of amino acids

Alanine Cysteine Arginine Alanine

Note that both the initial GCT and the final GCG code for alanine; because of the degenerate coding mentioned above, there are four codes for alanine in the standard genetic code.

[14] In fact, some cells can use the ratio of one alternative splicing to another to govern cellular behavior.

[15] Modifications to the RNA include the addition of a cap at the $5'$ end and a tail made of repeated adenine nucleotides (the poly-A tail).

[16] A complex containing hundreds of proteins and special-function RNA molecules.

[17] There are notable exceptions: the code for the naturally occurring amino acid selenocysteine is identical to that for a stop codon, except for a particular nucleotide sequence further downstream.

Most mRNA has a terminating poly-A tail. This terminal sequence makes it easy to pick out the labeling reactions that are used before hybridization to DNA microarrays, described in section 3.1.

Processing of amino acid chains Once the protein is formed, it has to find the right place to perform its function, whether as a structural protein in the cytoskeleton, or as a cell membrane receptor, or as a hormone that is to be secreted by the cell. There is a complex cellular apparatus that determines this translocation process. One of the determinants of the location and handling of a polypeptide is a portion of the polypeptide called the *signal peptide*. This header of amino acids is recognized by the translocation machinery and directs the ribosomal-mRNA complex to continue translation in a specific subcellular location, *e.g.*, constructing and inserting a protein into the endoplasmic reticulum for further processing and secretion by the cell. Alternatively, particular proteins may be delivered after translation and chaperones can prevent proper folding until the protein reaches its correct destination.

Transcriptional programs Initiation of the transcription process can be caused by external events or by a programmed event within the cell. For instance, the piezoelectric forces generated in bones through walking can gradually stimulate osteoblastic and osteoclastic transcriptional activity to cause bone remodeling. Similarly, heat shock or stress to the cell can cause rapid change or initiation of the transcriptional program. Additionally, changes in the microenvironment around the cell, such as the appearance of new micro- or macronutrients or the disappearance of these, will cause changes in the transcriptional program. Hormones secreted from distant organs bind to receptors which then directly or indirectly trigger a change in the transcriptional process.

There are also fully autonomous, internally programmed sequences of transcriptional expression. A classic example of this is the internal pacemaker that has been found with the *clock* and *per* genes (see figure 1.8 below) where, in the absence of any external stimuli, there is a recurring periodic pattern of transcriptional activity. Although this rhythmic pattern of transcription can be altered by external stimuli, nonetheless it will continue initiating this pattern of transcription without any additional stimuli.

Finally, there are pathological internal derangements of the cell which can lead to transcriptional activity. Self-repair or damage-detection programs may be internal to the cell, and can trigger self-destruction (called apoptosis) under certain conditions, such as irreparable DNA damage. As another example, there may be a deletion mutation of a repressor gene causing the gene normally repressed to in-

stead be highly active. There are many clinical instances of such disorders, such as familial male precocious puberty [161] where puberty starts at infancy due to a mutation in the luteinizing hormone receptor. This receptor normally activates only when luteinizing hormone is bound, but with the mutation present, activation does not require binding.

1.5.1 Biological *caveats* in mRNA measurements

There are several desiderata that should be understood about transcription in order to better understand the limitations of the gene profiling techniques.

- *The set of protein-coding RNA, otherwise known as the* transcriptome*, should be viewed as a pool of mRNA.* Even at equilibrium, each type of mRNA is degraded at its specific rate. New transcription provides additional mRNA to replace those being degraded, maintaining the levels. Thus, a transcriptional program should more accurately be viewed as a change in the transcriptome. In addition, just because a particular mRNA is seen in the transcriptome, it does not necessarily imply that the mRNA is about to be translated. Similarly, presence of a particular mRNA in the transcriptome does not mean that mRNA was recently transcribed. To determine what genes may be in the process of being translated, one can specifically look at polyribosomal mRNA, or that mRNA found in the presence of ribosomes. To determine which genes are currently being transcribed and processed, one can construct studies looking for pre-mRNA levels instead of mRNA levels (*i.e.*, by finding intronic splice sites, *etc.*).

- *mRNA can be transcribed at up to several hundred nucleotides per minute but may be transcribed at much slower rates.* In eukaryotic organisms, genes can take many hours to transcribe. For instance, a gene found in muscle, dystrophin, can take up to 20 hours to be transcribed. The lifespan of an mRNA molecule from initiation of transcription to ultimate degradation is a complex function of the rate of transcription, the stability of the particular mRNA molecule, and changes in its processing due to other cellular events, whether internally or externally initiated. Gene expression measurements, as performed by microarrays, measure the concentration of each mRNA as a snapshot at a single point of time or relative to another sample of mRNA. If the transcription of two genes is simultaneously stimulated and proceeds at the same rate, but the mRNA samples have different lifespans, then the expression measurements (and downstream protein production) from these two genes can be quite divergent.

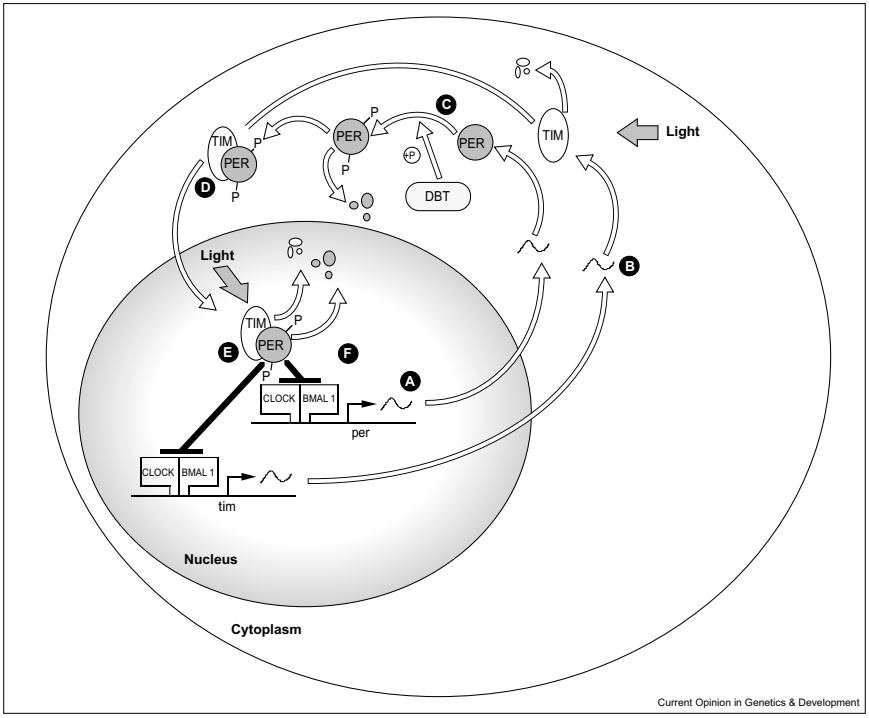

Figure 1.8
Genetic machinery of the circadian rhythm. Current molecular model of rhythm generation in *Drosophila*, from [191]. The succession of events (A–F) occur over the course of approximately 24 hours. *A*, CLOCKBMAL heterodimers bind the *per* and *tim* promoters and activate mRNA expression from each locus; CLOCKBMAL may also activate transcription of other circadian-regulated genes (not shown). *B*, *per* and *tim* mRNA are transported to the cytoplasm and translated into PER and TIM protein, respectively. *C*, Regulation of protein levels occurs by two mechanisms: DBT protein phosphorylates and destabilizes PER, and light destroys TIM. Light during the early subjective night can phase-delay the clock. Small "blobs" indicate degraded proteins. *D*, PER and TIM levels slowly accumulate during the early subjective night; TIM stabilizes PER and promotes nuclear transport. *E*, PER and TIM dimers enter the nucleus and inhibit CLOCKBMAL-activated transcription. *F*, Protein turnover (combined with the lack of new PER and TIM synthesis) leads to derepression of *per* and *tim* mRNA expression; the cycle begins again (*A*). Light during the late subjective night can phase-advance the clock.

- *Proteins perform most cellular functions.* The lifespan of proteins is at least as variable as that of mRNA. Consequently, measurements of gene expression (*i.e.*, mRNA measurement) may not accurately correspond to the concentration or activity of the protein for which it codes. This should be a caution to any investigator imputing function to a gene based solely on gene expression patterns.

- *The mRNA generated from the transcription of a gene may differ depending on which exons and introns are or are not spliced into the mRNA molecule.* These alternate splicing products are mRNA molecules with divergent sequences. Therefore, measurements of mRNA that are designed to measure only one of these alternate splicing products will provide incomplete if not misleading information.

- *The genetic basis for organismal diversity is due in large part to differences in sequences, also known as polymorphisms, of each gene.* Most of these polymorphisms differ from one another by one nucleotide and are known as single nucleotide polymorphisms (SNPs). Due to the small portion of the genome coding for proteins and to redundancy in the mRNA code, described on page 27, only some SNPs will result in differently constructed proteins. If a gene expression measurement technology is highly specific for a particular SNP, then other variants will not be measured. When we consider that the Human Genome Project to date only includes the sequence of handfuls of individuals, the implications of such measurement specificity become apparent. Nonetheless, as common SNPs become documented for each gene, it is likely that successful expression measurement techniques will measure each of them.

Despite these *caveats*, it remains that gene profiling studies have proven to be remarkably robust in describing the functional grouping and coordinated behavior of genes. The reasons for this are discussed in section 2.2, page 60.

1.5.2 Sequence-level genomics

Sequencing or genotyping the genome of individuals allows the characterization of what distinguishes the hereditable material of each individual from that of others. By matching differences in phenotype (*e.g.*, blood pressure, adult height) between individuals and this genomic characterization (*i.e.*, in association studies) genetic epidemiologists are able to impute these phenotypic differences to a small span of the genome. The smallness of the span is a function of the spatial resolution with which the genomic characterization occurs. Although there is controversy around what constitutes sufficient resolution, there is some consensus that genomic markers such as SNPs spaced every thousand bases will be sufficient to unambiguously resolve

the span of the genome associated with a phenotypic difference to a single gene [112]. Currently, the cost of a single genotype is around $0.50 and so the cost of a high-resolution genome scan of an individual is on the order of magnitude of $1 million. At this cost, only a very few institutions can afford a comprehensive study of a population. If the recent past is to be a guide, the cost of genotyping is likely to drop by several orders of magnitude well within a decade, at which point genome-wide scans of populations will become economically feasible.

The kind of information that these studies will provide includes the contribution of particular polymorphisms to changes in phenotype, presumably via changes in gene function, pattern of expression, or both. Currently, expression microarrays capture information about gene polymorphism poorly, if at all. Presently shipping microarrays typically code for a canonical "normal" gene sequence. Departures from this sequence will result in changes in the intensity reading reported by any of the currently employed microarray platforms, as will become obvious after reading the chapter on measurement techniques (chapter 3). Undoubtedly, in the next 5 years, the continued geometric increase in the complexity and density of expression microarrays will allow specific assaying for all common and all clinically significant polymorphisms. Until then, however, polymorphisms will be a source of "noise" or unexplained variation. Functional genomics investigations designed to elucidate the role of particular polymorphisms in phenotypic mechanisms or phenotypic variation are best conducted using methods other than RNA expression detection microarrays, such as high-throughput sequencing and genotyping. These latter techniques are beyond the scope of this text (see [43, 112] for an excellent survey of sequencing and genotyping) and therefore this particular avenue of functional genomics will not be addressed below except tangentially.

1.5.3 Proteomics

There is a good deal of enthusiasm at present about the emerging discipline of proteomics. The promise of proteomics is that we will be able to measure, in a similarly comprehensive and parallel fashion to RNA microarray measurements, the concentrations of proteins present in a particular cellular system. In a very abstract sense, for the purely computational bioinformatician, the field of proteomics does not harbor any particular novelty in that all it provides is another 100,000 variables, or more, to describe the state of cellular processes. In that perspective, a proteomics data set reduces simply to another array that has distinguishing noise characteristics (*i.e.*, sources of biological and measurement variability as described in section 3.2.5) and is just as amenable to the clustering and classification techniques of chapter 4 as any set of microarray expression measurements. Less abstractly, how-

ever, proteomics offers a set of insights that are quite different and divergent from those of expression microarrays. The assumption underlying expression microarray measurements is that by capturing the patterns of expression management, we will capture the basic irregulatory rhythms of the cell [34]. Although these assumptions may hold at times and have done remarkably well in helping biologists elucidate some fundamental biology and to classify clinical phenomena, there are several persuasive reasons why these assumptions should not always hold. First, we know that most of the effector molecules in cellular metabolism are proteins. To the extent that the timing of protein synthesis and the half-life of proteins is not closely coupled to that of RNA expression, the assumption of the representativeness of RNA levels does not hold. As outlined in section 1.5, these assumptions do not hold in many instances. Nonetheless, in proteomics, we will be bedeviled by a new set of assumptions that will be equally problematic and challenging. Assuming that we have high reproducibility and compact systems for assessing the concentration of tens of thousands of proteins, we will be faced with the following challenges:

- Similar concentrations do not imply co-regulation. Given that proteins have hugely different half-lives, even within a single cell (*e.g.*, a structural protein in a bone osteoblast and a parathyroid hormone receptor in the same cell), then the concentrations of protein molecules in a cell may only remotely reflect joint regulation. This problem also haunts the analysis of RNA expression microarray data because of the wide range of the stability-degradation rate of mRNA.

- Conversely, repeatedly different concentrations of two proteins imply co-regulation. At any given sampling time, the two proteins could have quite variable concentrations and different mutual relationships. Yet, there is nothing about this to preclude important functional interactions between these proteins.[18]

- Localization heterogeneity. Unlike transcription of genes, which occurs within the nucleus, protein activity has very distinct and heterogeneous functional significance in different parts of the cellular compartments, and therefore an essential part of understanding protein function and regulation from proteomic data will require detailed localization to subcompartments of organelles in order to be meaningful. This problem also exists with RNA expression microarray data, because of the differing biological implications of RNA concentrations measured before splicing, after splicing in the cytoplasm, and during translation.

[18] One way to circumvent this problem is with multiple serial measurements of all proteins and then to subject them to a dynamics analysis.

These challenges of proteomics will eventually be addressed by novel ways of looking at protein activity over time and in different spatial locations. Nonetheless, at present, the basic mechanism for cheaply and reliably obtaining large numbers of parallel measurements of protein activity have yet to be worked out and industrialized, so that these developments are not likely to occur on a large scale for at least 1 or 2 years. When these challenges have been resolved, then indeed the arrays of proteinomic data will be amenable to the same techniques of analysis as described in this book for RNA expression. For thoughtful and comprehensive insights into the challenges of proteomics and data analytic techniques, we refer the reader to the following papers [25, 69, 141] and websites `http://www.expasy.ch/`, `http://www.hip.harvard.edu/`.

References

[1] T. Akutsu, S. Miyano, and S. Kuhara. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of computational biology*, 7(3-4):331–43, 2000.

[2] T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–34, 2000.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, 2000.

[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. 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 of the United States of America*, 96(12):6745–50, 1999.

[5] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):10101–6, 2000.

[6] R. B. Altman and S. Raychaudhuri. Whole-genome expression analysis: challenges beyond clustering. *Current opinion in structural biology*, 11(3):340–7, 2001.

[7] Anonymous. Jargon file version 4.0.0, 1996.

[8] Anonymous. Array data go public. *Nature genetics*, 22(3):211–2, 1999.

[9] Anonymous. GEM microarray reproducibility study. Technical Report PN 99-0169, Incyte Pharmaceuticals, Inc., September 1999.

[10] Anonymous. *GeneChip analysis suite user guide*, volume version 3.3. Affymetrix, Inc., 1999.

[11] S. M. Arfin, A. D. Long, E. T. Ito, L. Tolleri, M. M. Riehle, E. S. Paegle, and G. W. Hatfield. Global gene expression profiling in Escherichia coli K12. The effects of integration host factor. *The Journal of biological chemistry*, 275(38):29672–84, 2000.

[12] A. Arkin, P. Shen, and J. Ross. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277:1275–9, 1997.

[13] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–9, 2000.

[14] Roger Bacon. *Opus majus*. Russell and Russell, New York, 1962.

[15] J. Banfield and A. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[16] D. E. Bassett, Jr., M. B. Eisen, and M. S. Boguski. Gene expression informatics–it's all in your mine. *Nature genetics*, 21(1 Suppl):51–5, 1999.

[17] L. R. Baugh, A. A. Hill, E. L. Brown, and C. P. Hunter. Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic acids research*, 29(5):E29, 2001.

[18] C. F. Belanger, C. H. Hennekens, B. Rosner, and F. E. Speizer. The nurses' health study. *The American journal of nursing*, 78(6):1039–40, 1978.

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

[20] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–97, 1999.

[21] Amir Ben-Dor, Nir Friedman, and Zohar Yakhini. Tissue classification with gene expression profiles. In *RECOMB*, pages 31–38, Tokyo, Japan, 1999. ACM.

[22] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. GenBank. *Nucleic acids research*, 28(1):15–8, 2000.

[23] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of computational biology*, 5(1):27–40, 1998.

[24] F. Bertucci, K. Bernard, B. Loriod, Y. C. Chang, S. Granjeaud, D. Birnbaum, C. Nguyen, K. Peck, and B. R. Jordan. Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. *Human molecular genetics*, 8(9):1715–22, 1999.

[25] V. E. Bichsel, L. A. Liotta, and E. F. Petricoin, 3rd. Cancer proteomics: from biomarker discovery to signal pathway profiling. *Cancer journal*, 7(1):69–78, 2001.

[26] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–40, 2000.

[27] M. Bittner, P. Meltzer, and J. Trent. Data analysis and integration: of steps and arrows. *Nature genetics*, 22(3):213–5, 1999.

[28] M. S. Boguski and G. D. Schuler. ESTablishing a human transcript map. *Nature genetics*, 10(4):369–71, 1995.

[29] D. D. Bowtell. Options available–from start to finish–for obtaining expression data by microarray. *Nature genetics*, 21(1 Suppl):25–32, 1999.

[30] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS letters*, 480(1):17–24, 2000.

[31] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630–4, 2000.

[32] B. M. Broccolo and B. W. Petersen. Final HIPAA privacy rules: "How do we get started?". *Journal of health care finance*, 27(4):7–23, 2001.

[33] Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, Jr., and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262–267, 2000.

[34] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature genetics*, 21(1 Suppl):33–7, 1999.

[35] Michael Brush. Making sense of microchip array data. *The Scientist*, 15(9):25, 2001.

[36] W. Buntine. Operations for learning with graphical models. *Journal of artificial intelligence research*, 2:159–225, 1994.

[37] A. J. Butte and I. S. Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. In Nancy Lorenzi, editor, *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 711–715, Washington, DC, 1999. Hanley & Belfus.

[38] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 5, pages 418–29, Hawaii, 2000.

[39] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):12182–6, 2000.

[40] Atul J. Butte, Jessica Ye, G. Niederfellner, K. Rett, H. U. Hädring, Morris F. White, and Issac S. Kohane. Determining significant fold differences in gene expression analysis. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 6, pages 6–17, Hawaii, 2001.

[41] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M. C. Hermus, R. van Asperen, K. Boon, P. A. Voute, S. Heisterkamp, A. van Kampen, and R. Versteeg. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 291(5507):1289–92, 2001.

[42] G. Casella and R. L. Berger. *Statistical inference*. Statistics and Probability. Brooks/Cole Publishing Company, Pacific Grove, CA, 1990.

[43] A. Chakravarti. Population genetics–making sense out of sequence. *Nature genetics*, 21(1 Suppl):56–60, 1999.

[44] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 4, pages 29–40, Hawaii, 1999.

[45] Y. Chen, E. Dougherty, and M. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of biomedical optics*, 2(4):364–374, 1997.

[46] M. L. Chow, E. J. Moler, and I. S. Mian. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiological genomics*, 5(2):99–111, 2001.

[47] Paul D. Clayton, W. Earl Boebert, Gordon H. Defriese, Susan P. Dowell, Mary L. Fennell, Kathleen A. Frawley, John Glaser, Richard A. Kemmerer, Carl E. Landwehr, Thomas C. Rindfleisch, Sheila A. Ryan, Bruce J. Sams, Jr., Peter Szolovits, Robbie G. Trussell, and Elizabeth Ward. *For the record: protecting electronic health information*. National Academy Press, Washington, DC, 1997.

[48] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[49] E. Coiera. Clinical communication: a new informatics paradigm. In James Cimino, editor, *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 17–21, Washington, DC, 1996. Hanley & Belfus.

[50] E. Coiera and V. Tombs. Communication behaviours in a hospital setting: an observational study. *BMJ*, 316(7132):673–6, 1998.

[51] M. H. Coletti and H. L. Bleich. Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4):317–323, 2001.

[52] G. F. Cooper and E. Herskovitz. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9:309–347, 1992.

[53] R. A. Cote and S. Robboy. Progress in medical information management: systematized nomenclature of medicine (SNOMED). *Journal of the American Medical Association*, 243:756–762, 1980.

[54] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, 1991.

[55] T. R. Dawber, G. F. Meadors, and F. E. J. Moore. The Framingham study: Epidemiological approaches to heart disease. *American journal of public health*, 41:279–286, 1951.

[56] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature genetics*, 14(4):457–60, 1996.

[57] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–6, 1997.

[58] R. L. Devaney. *An introduction to chaotic dynamical systems*. Addison-Wesley, Redwood City, CA, 1989.

[59] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[60] Joaquin Dopazo, Edward Zanders, Ilaria Dragoni, Gillian Amphlett, and Francesco Falciani. Methods and approaches in the analysis of gene expression data. *Journal of immunological methods*, 250(1-2):93–112, 2001.

[61] Sandrine Dudoit. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, University of California, Berkeley, August 2000.

[62] B. L. Ebert and H. F. Bunn. Regulation of transcription by hypoxia requires a multiprotein complex that includes hypoxia-inducible factor 1, an adjacent transcription factor, and p300/CREB binding protein. *Molecular and cellular biology*, 18(7):4089–96, 1998.

[63] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–8, 1998.

[64] D. Endy and R. Brent. Modelling cellular behaviour. *Nature*, 409 Suppl(6818):391–5, 2001.

[65] O. Ermolaeva, M. Rastogi, K. D. Pruitt, G. D. Schuler, M. L. Bittner, Y. Chen, R. Simon, P. Meltzer, J. M. Trent, and M. S. Boguski. Data management and analysis for gene expression arrays. *Nature genetics*, 20(1):19–23, 1998.

[66] R. M. Ewing, A. B. Kahla, O. Poirot, F. Lopez, S. Audic, and J. M. Claverie. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome research*, 9(10):950–9, 1999.

[67] L. A. Farrer, L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux, R. H. Myers, M. A. Pericak-Vance, N. Risch, and C. M. van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. *Journal of the American Medical Association*, 278(16):1349–56, 1997.

[68] O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R. N. Trethewey, and L. Willmitzer. Metabolite profiling for plant functional genomics. *Nature biotechnology*, 18(11):1157–61, 2000.

[69] D. Figeys and D. Pinto. Proteomics on a chip: promising developments. *Electrophoresis*, 22(2):208–16, 2001.

[70] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–73, 1991.

[71] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–20, 2000.

[72] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–14, 2000.

[73] T. Gaasterland and S. Bekiranov. Making the most of microarray data. *Nature genetics*, 24(3):204–6, 2000.

[74] F. Gamarra, G. Simic-Schleicher, R. M. Huber, A. Ulsenheimer, P. C. Scriba, U. Kuhnle, and M. Wehling. Impaired rapid mineralocorticoid action on free intracellular calcium in pseudohypoaldosteronism. *The Journal of clinical endocrinology and metabolism*, 82(3):831–4, 1997.

[75] G. K. Geiss, R. E. Bumgarner, M. C. An, M. B. Agy, A. B. van 't Wout, E. Hammersmark, V. S. Carter, D. Upchurch, J. I. Mullins, and M. G. Katze. Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology*, 266(1):8–16, 2000.

[76] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):12079–84, 2000.

[77] W. Wayt Gibbs. Shrinking to enormity. *Scientific American*, 284(2):33–34, February 2001.

[78] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, 1999.

[79] S. Greenfield, S. Cretin, L. G. Worthman, and F. Dorey. The use of an ROC curve to express quality of care results. *Medical decision making*, 2(1):23–31, 1982.

[80] J. Guckenheimer and P. Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1983.

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

[82] Ira J. Haimowitz, Ramesh S. Patil, and Peter Szolovits. Representing medical knowledge in a terminological language is difficult. In R. A. Greenes, editor, *Proceedings of the Twelfth Symposium on Computer Applications in Medical Care*, pages 101–105, Washington, DC, 1988. IEEE Computer Society Press.

[83] J. Hanke and J. G. Reich. Kohonen map as a visualization tool for the analysis of protein sequences: multiple alignments, domains and segments of secondary structures. *Computer applications in the biosciences*, 12(6):447–54, 1996.

[84] A. Harding and C. Stuart-Buttle. The development and role of the Read Codes. *Journal of American Health Information Management Association*, 69(5):34–8, 1998.

[85] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 6, pages 422–33, Hawaii, 2001.

[86] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown. Supervised harvesting of expression trees. *Genome biology*, 2(1), 2001.

[87] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown. "gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome biology*, 1(2), 2000.

[88] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. Technical Report MSR-TR-94-09, Microsoft Research, March 1994.

[89] David Heckerman. Bayesian networks for data mining. *Data mining and knowledge discovery*, 1(1):79–119, 1997.

[90] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene-expression profiles in hereditary breast cancer. *The New England journal of medicine*, 344(8):539–48, 2001.

[91] A. Herbert and A. Rich. RNA processing and the evolution of eukaryotes. *Nature genetics*, 21(3):265–9, 1999.

[92] J. Herrero, A. Valencia, and J. Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126–36, 2001.

[93] Ralf Herwig, Albert J. Poustka, Christine Muller, Christof Bull, Hans Lehrach, and John O'Brien. Large-scale clustering of cDNA-fingerprinting data. *Genome research*, 9:1093–1105, 1999.

[94] L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome research*, 9(11):1106–15, 1999.

[95] S. G. Hilsenbeck, W. E. Friedrichs, R. Schiff, P. O'Connell, R. K. Hansen, C. K. Osborne, and S. A. Fuqua. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *Journal of the National Cancer Institute*, 91(5):453–9, 1999.

[96] K. Hokamp and K. Wolfe. What's new in the library? What's new in GenBank? Let PubCrawler tell you. *Trends in genetics*, 15(11):471–2, 1999.

[97] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–28, 1998.

[98] Yuh-Jyh Hu. An integrated approach for genome-wide gene expression analysis. *Computer methods and programs in biomedicine*, 65(3):163–174, 2001.

[99] S. M. Huff, R. A. Rocha, C. J. McDonald, G. J. De Moor, T. Fiers, W. D. Bidgood, Jr., A. W. Forrey, W. G. Francis, W. R. Tracy, D. Leavelle, F. Stalling, B. Griffin, P. Maloney, D. Leland, L. Charles, K. Hutchins, and J. Baenziger. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. *Journal of the American Medical Informatics Association*, 5(3):276–92, 1998.

[100] M. Ishii, S. Hashimoto, S. Tsutsumi, Y. Wada, K. Matsushima, T. Kodama, and H. Aburatani. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, 68(2):136–43, 2000.

[101] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[102] K. K. Jain. Tech.Sight. Biochips for gene spotting. *Science*, 294(5542):621–3, 2001.

[103] M. Janowitz. A relational approach to ordinal clustering and classification. 2000. In preparation.

[104] R. A. Jungmann, D. Huang, and D. Tian. Regulation of LDH-A gene expression by transcriptional and posttranscriptional signal transduction mechanisms. *The Journal of experimental zoology*, 282(1-2):188–95, 1998.

[105] R. E. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[106] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of computational biology*, 7(6):819–37, 2000.

[107] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–9, 2001.

[108] Ju Han Kim, Lucila Ohno-Machado, and Isaac S. Kohane. Unsupervised learning from complex data: the matrix incision tree algorithm. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 6, pages 30–41, Hawaii, 2001.

[109] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics*. Springer-Verlag, New York, 1992.

[110] G. T. Klus, A. Song, A. Schick, M. Wahde, and Z. Szallasi. Mutual information analysis as a tool to assess the role of aneuploidy in the generation of cancer-associated differential gene expression patterns. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 6, pages 42–51, Hawaii, 2001.

[111] Isaac S. Kohane, Hongmei Dong, and Peter Szolovits. Health information identification and de-identification toolkit. In Christopher Chute, editor, *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 356–360, Philadelphia, PA, 1998. Hanley & Belfus.

[112] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature genetics*, 22(2):139–44, 1999.

[113] Winston P. Kuo, Tor-Kristian Jenssen, Atul J. Butte, L. Ohno-Machado, and Isaac S. Kohane. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, 18(3):405–412, 2002.

[114] R. H. Lathrop. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein engineering*, 7(9):1059–68, 1994.

[115] S. L. Lauritzen. *Graphical models*. Oxford University Press, New York, 1996.

[116] C. K. Lee, R. G. Klopp, R. Weindruch, and T. A. Prolla. Gene expression profile of aging and its retardation by caloric restriction. *Science*, 285(5432):1390–3, 1999.

[117] C. K. Lee, R. Weindruch, and T. A. Prolla. Gene-expression profile of the ageing brain in mice. *Nature genetics*, 25(3):294–297, 2000.

[118] G. Lennon, C. Auffray, M. Polymeropoulos, and M. B. Soares. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics*, 33(1):151–2, 1996.

[119] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 3, pages 18–29, Hawaii, 1998.

[120] D. A. B. Lindberg and B. L. Humphreys. The Unified Medical Language System (UMLS) and computer-based patient records. In M. J. Ball and M. F. Collen, editors, *Aspects of the computer-based patient record*, pages 165–175. Springer-Verlag, New York, 1992.

[121] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nature genetics*, 21(1 Suppl):20–4, 1999.

[122] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675–80, 1996.

[123] L. Luo, R. C. Salunga, H. Guo, A. Bittner, K. C. Joy, J. E. Galindo, H. Xiao, K. E. Rogers, J. S. Wan, M. R. Jackson, and M. G. Erlander. Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature medicine*, 5(1):117–22, 1999.

[124] S. L. Madden, C. J. Wang, and G. Landes. Serial analysis of gene expression: from gene discovery to target identification. *Drug discovery today*, 5(9):415–425, 2000.

[125] K. D. Mandl, P. Szolovits, and I. S. Kohane. Public standards and patients' control: how to keep electronic medical records accessible but private. *BMJ*, 322(7281):283–7, 2001.

[126] P. M. Mannucci. Polymorphisms in the factor VII gene and the risk of myocardial infarction. *The New England journal of medicine*, 344(6):458–9, 2001.

[127] D. R. Masys, J. B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319–26, 2001.

[128] H. Matsuno, A. Doi, M. Nagasaki, and S. Miyano. Hybrid petri net representation of gene regulatory network. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 5, pages 341–52, Hawaii, 2000.

[129] D. B. McCarn. MEDLINE: An introduction to on-line searching. *Journal of the American Society for Information Science*, 31(3):181–192, 1980.

[130] Clement J. McDonald, L. Blevins, W. M. Tierney, and D .K. Martin. The Regenstrief medical records. *MD Computing*, 5(5):34–47, 1988.

[131] R. McEntire, P. Karp, N. Abernethy, D. Benton, G. Helt, M. DeJongh, R. Kent, A. Kosky, S. Lewis, D. Hodnett, E. Neumann, F. Olken, D. Pathak, P. Tarczy-Hornoch, L. Toldo, and T. Topaloglou. An evaluation of ontology exchange languages for bioinformatics. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, volume 8, pages 239–50, San Diego, 2000.

[132] A. A. Mironov, J. W. Fickett, and M. S. Gelfand. Frequent alternative splicing of human genes. *Genome research*, 9(12):1288–93, 1999.

[133] E. J. Moler, D. C. Radisky, and I. S. Mian. Integrating naive bayes models and external knowledge to examine copper and iron homeostasis in S. cerevisiae. *Physiological genomics*, 4(2):127–135, 2000.

[134] Kevin Murphy and Saira Mian. Modelling gene expression data using dynamic bayesian networks. Technical report, Lawrence Berkeley National Laboratory, 1999.

[135] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational biology*, 8(1):37–52, 2001.

[136] A. Nimgaonkar, D. Sanoudou, A. J. Butte, J. N. Haslett, L. M. Kunkel, A. H. Beggs, and I. S. Kohane. Reproducibility of gene expression across generations of Affymetrix microarrays. 2002. Submitted.

[137] R. Patil, R. Fikes, P. Patel-Schneider, D. McKay, T. Finin, T. Gruber, and R. Neches. The DARPA knowledge sharing effort: Progress report. In *Principles of Knowledge Representation and Reasoning: Third International Conference*, Royal Sonesta Hotel, Cambridge, MA, 1992.

[138] M. E. Patti, X. J. Sun, J. C. Bruening, E. Araki, M. A. Lipes, M. F. White, and C. R. Kahn. 4PS/insulin receptor substrate (IRS)-2 is the alternative substrate of the insulin receptor in IRS-1-deficient mice. *The Journal of biological chemistry*, 270(42):24670–3, 1995.

[139] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[140] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

[141] A. Persidis. Proteomics. *Nature biotechnology*, 16(4):393–4, 1998.

[142] D. M. Pisanelli, A. Gangemi, and G. Steve. WWW-available conceptual integration of medical terminologies: the ONIONS experience. In Daniel Masys, editor, *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 575–579, Nashville, TN, 1997. Hanley & Belfus.

[143] K. R. Popper. *The logic of scientific discovery*. Basic Books, New York, 1959.

[144] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes: the art of scientific computing*. Cambridge University Press, New York, 2nd edition, 1992.

[145] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic acids research*, 29(1):137–40, 2001.

[146] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[147] J. Quinn. An HL7 (Health Level Seven) overview. *Journal of American Health Information Management Association*, 70(7):32–4, 1999.

[148] M. Ramoni and P. Sebastiani. Bayesian methods. In M. Berthold and D. J. Hand, editors, *Intelligent Data Analysis. An Introduction*, pages 129–166. Springer, New York, NY, 1999.

[149] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 5, pages 455–66, Hawaii, 2000.

[150] B. Y. Reis, A. Butte, and I. S. Kohane. Extracting knowledge from dynamics in gene expression. *Journal of biomedical informatics*, 34(1):15–27, 2001.

[151] K. Reue. mRNA quantitation techniques: considerations for experimental design and application. *The Journal of nutrition*, 128(11):2038–44, 1998.

[152] C. S. Richmond, J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner. Genome-wide expression profiling in Escherichia coli K-12. *Nucleic acids research*, 27(19):3821–3835, 1999.

[153] David M. Rind, Isaac S. Kohane, Peter Szolovits, Charles Safran, Henry C. Chueh, and G. Octo Barnett. Maintaining the confidentiality of medical records shared over the internet and world wide web. *Annals of internal medicine*, 127(2):138–141, 1997.

[154] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T.G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics*, 24(3):227–35, 2000.

[155] D. E. Rumelhart and J. L. McClelland. *Parallel distributed processing: explorations in the microstructure of cognition*. MIT Press, Cambridge, MA, 1986.

[156] E. E. Schadt, C. Li, C. Su, and W. H. Wong. Analyzing high-density oligonucleotide gene expression array data. *Journal of cellular biochemistry*, 80(2):192–202, 2000.

[157] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70, 1995.

[158] P. L. Schuyler, W. T. Hole, M. S. Tuttle, and D. D. Sherertz. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217–22, 1993.

[159] J. A. Segal, J. L. Barnett, and D. L. Crawford. Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. *Journal of molecular evolution*, 49(6):736–49, 1999.

[160] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:623–58, 1948.

[161] A. Shenker, L. Laue, S. Kosugi, J. J. Merendino, Jr., T. Minegishi, and G. B. Cutler, Jr. A constitutively activating mutation of the luteinizing hormone receptor in familial male precocious puberty. *Nature*, 365(6447):652–4, 1993.

[162] G. Sherlock. Analysis of large-scale gene expression data. *Current opinion in immunology*, 12(2):201–5, 2000.

[163] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, Cambridge, MA, 2nd edition, 1996.

[164] R. Somogyi and C. A. Sniegoski. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity*, 1(6):45–63, 1996.

[165] A. Soukas, P. Cohen, N. D. Socci, and J. M. Friedman. Leptin-specific patterns of gene expression in white adipose tissue. *Genes & development*, 14(8):963–80, 2000.

[166] E. Southern, K. Mir, and M. Shchepinov. Molecular interactions on microarrays. *Nature genetics*, 21(1 Suppl):5–9, 1999.

[167] 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(12):3273–97, 1998.

[168] D. J. Spiegelhalter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:157–224, 1990.

[169] Latanya Sweeney. Replacing personally-identifying information in medical records, the SCRUB system. In James Cimino, editor, *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 333–337, Washington, DC, 1996. Hanley & Belfus.

[170] Latanya Sweeney. Guaranteeing anonymity when sharing medical data, the Datafly system. In Daniel Masys, editor, *Proceedings of the American Medical Informatics Association Fall Symposium*, pages 51–55, Nashville, TN, 1997. Hanley & Belfus.

[171] Latanya Sweeney. Three computational systems for disclosing medical data in the year 1999. In B. Cesnik, A. T. McCray, and J. R. Scherrer, editors, *Medinfo*, volume 9 Pt 2, pages 1124–9, Seoul, Korea, 1998. IOS Press.

[172] Z. Szallasi and S. Liang. Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 3, pages 66–76, Hawaii, 1998.

[173] P. Szolovits and S. G. Pauker. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11:115–144, 1978.

[174] Peter Szolovits and Isaac Kohane. Against simple universal health identifiers. *Journal of the American Medical Informatics Association*, 1(4):316–319, 1994.

[175] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):2907–12, 1999.

[176] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281–5, 1999.

[177] D. Thieffry and R. Thomas. Qualitative analysis of gene networks. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 3, pages 77–88, Hawaii, 1998.

[178] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a dataset via the gap statistic. Technical report, Department of Statistics, Stanford University, March 2000.

[179] M. Tomita, K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. A. Hutchison, III. E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15(1):72–84, 1999.

[180] P. Toronen, M. Kolehmainen, G. Wong, and E. Castren. Analysis of gene expression data using self-organizing maps. *FEBS letters*, 451(2):142–6, 1999.

[181] C. L. Tsien, T. A. Libermann, X. Gu, and I. S. Kohane. On reporting fold differences. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 6, pages 496–507, Hawaii, 2001.

[182] E. P. van Someren, L. F. Wessels, and M. J. Reinders. Linear modeling of genetic networks from experimental data. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, volume 8, pages 355–66, 2000.

[183] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995.

[184] V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler. Characterization of the yeast transcriptome. *Cell*, 88(2):243–51, 1997.

[185] M. Wahde and J. Hertz. Modeling genetic regulatory dynamics in neural development. *Journal of computational biology*, 8(4):429–42, 2001.

[186] D. C. Weaver, C. T. Workman, and G. D. Stormo. Modeling regulatory networks with weight matrices. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 4, pages 112–23, Hawaii, 1999.

[187] M. C. Weinstein, H. V. Fineberg, A. S. Elstein, H. S. Frazier, D. Neuhauser, R. R. Neutra, and B. J. McNeil. *Clinical decision analysis*. W. B. Saunders, Philadelphia, 1980.

[188] Sholom M. Weiss and Nitin Indurkhya. *Predictive data mining: a practical guide*. Morgan Kaufmann Publishers, San Francisco, 1997.

[189] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America*, 95(1):334–9, 1998.

[190] J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley, New York, 1990.

[191] L. D. Wilsbacher and J. S. Takahashi. Circadian rhythms: molecular basis of the clock. *Current opinion in genetics & development*, 8(5):595–602, 1998.

[192] D. J. Withers, D. J. Burks, H. H. Towery, S. L. Altamuro, C. L. Flint, and M. F. White. Irs-2 coordinates Igf-1 receptor-mediated beta-cell development and peripheral insulin signalling. *Nature genetics*, 23(1):32–40, 1999.

[193] Peter J. Woolf and Yixin Wang. A fuzzy logic approach to analyzing gene expression data. *Physiological genomics*, 3(1):9–15, 2000.

[194] S. Wright. Correlation and causation. *Journal of agricultural research*, 20:557–585, 1921.

[195] S. Wright. The theory of path coefficients: a reply to Niles' criticism. *Genetics*, 8:239–255, 1923.

[196] S. Wright. The method of path coefficients. *Annals of mathematical statistics*, 5:161–215, 1934.

[197] A. Wuensche. Genomic regulation modeled as a network with basins of attraction. In R. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 3, pages 89–102, Hawaii, 1998.

[198] 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(6760):418–21, 1999.

[199] Yee Hwa Yang, Michael J. Buckley, Sandrine Dudoit, and Terence P. Speed. Comparison of methods for image analysis on cDNA microarray data. Technical Report 584, Univeristy of California, Berkeley, 2000.

[200] D. B. Young. A post-genomic perspective. *Nature medicine*, 7(1):11–3, 2001.