1

Introduction to Computational Approaches for Biology and Medicine

Sarah J. Aerni and Marina Sirota

Stanford University School of Medicine, Biomedical Informatics Training Program, Stanford, California 94305

The past decade has witnessed a staggering increase in the abundance and availability of molecular data produced, for example, from experiments in gene expression, proteomics, and DNA sequencing. With the expansion of data generation across these various areas of biomedicine, developing computational techniques and approaches has become instrumental not only in enhancing and advancing scientific discovery, but also in the analysis of these discoveries. However, many biomedical researchers are unaware of the ways in which informatics can improve their progress, finding it inaccessible because of what we term "informatics anxiety." We believe that this concern results from the improper communication of the intuition behind the approaches as well as the value of tools available.

The goal of this book is to enable biomedical researchers to use and develop computational tools by making these tools more understandable and accessible. To accomplish this goal, we approach the field from the viewpoint of a biomedical researcher: (1) arming the biologist with a basic understanding of the fundamental concepts in the field; (2) presenting tools from the standpoint of the data for which they are created; and (3) showing how the field of informatics is quickly adapting to the challenges and advances in biomedical technologies. After summarizing these points, we provide a series of brief descriptions for each chapter and a set of questions that the reader should be able to answer based on the material presented.

OBJECTIVES

Many wet-bench biomedical researchers experience *bioinformatics anxiety*. The various existing tools are often described in extensive mathematical detail to show the

rigor of a model to fellow bioinformaticians. However, in the absence of explanations that are more accessible to scientists without extensive training in the field, the tools often go unused. The intended end users are therefore unable to understand how or why this tool should be applied to their own data.

Many informatics tools, although designed for use by the biomedical researcher, rarely have the impact hoped for by the bioinformaticians. This is largely due to the inability to communicate the methods and utility of the tools generally in a way that a biomedical researcher can properly understand. Typically, the tools described in the literature are applied a single time on a benchmark data set that the bioinformatics researchers used in development. However, the goal of these papers is largely to show a proof-of-principle to the biomedical research community in general, with the intention of adoption of the method in the community as a standard tool. Yet, many biomedical researchers appear to use only the tools they know, finding the process of understanding and incorporating new methods into their analysis a daunting task they prefer not to tackle.

The purpose of this book is to arm molecular biologists with the knowledge to enable their use of bioinformatics tools. These tools will undeniably enhance their work and lead to important discoveries. Over the years, bioinformatics has enabled scientists to pose and ask questions that would not be possible otherwise. Many of these have yielded success in various fields of biomedicine such as novel drug discoveries (Gleevec in the treatment of some leukemias), genetic interactions with various drug and disease phenotypes (Warfarin, an anticoagulant used to prevent clotting), and drug repositioning (applying known drugs in new treatment scenarios).

ORGANIZATION

In this volume, we provide biomedical researchers with the principles that form the basis of many bioinformatics methods and examples of how these techniques can be used to analyze diverse data sets. The book is organized in three sections: First, we tackle the fundamental concepts underlying most bioinformatics methods. Next, we present a series of data-specific techniques available to biomedical researchers. We conclude with integrative techniques that augment the biomedical researcher's primary analysis with additional complementary data.

Fundamental Concepts for Bioinformatics Methods

This section serves as a reference for how the basic principles are used in bioinformatics approaches. Here, we discuss the underlying technical knowledge in the fundamental fields of bioinformatics in the fields of computer science, statistics, and machine learning. These very technical fields are presented at a level of detail that provides the necessary intuition for understanding the methods without requiring expert knowledge. With this level of detail, the reader will be able to understand

the inner workings of the tools that are currently available as well as new approaches beyond those described in the book.

- ***Chapter 2: Introduction to Computer Science.*** Chapter 2 covers the practicality of methods, algorithm development, and understanding what computational complexity means for the user. This chapter gives the framework for using and understanding existing published tools. What is a computer? How is information stored?
- ***Chapter 3: Probability and Statistics.*** This chapter presents the techniques from the field that reappear in many bioinformatics papers. How can the biomedical researcher assess whether the results obtained are significant? Which metric is appropriate for a given data set? These fundamentals will help biomedical researchers speak and understand the language of informatics. The chapter also discusses some basic and commonly used techniques to answer the following questions: What kind of correction needs to be applied to assess significance? How do I assess the quality of an informatics approach?
- ***Chapter 4: Machine Learning.*** Application of various machine learning techniques are presented in the context of biological data geared toward the noncomputationally trained life scientist. For various algorithms, the intuition is presented to help biologists understand when it is appropriate to use each one. What is the difference between a supervised and an unsupervised approach? Which algorithm should I use on my data?

Techniques for Analyzing Your Data

In the second section, we present a set of methods to help researchers understand what tools are available to perform an automated analysis on their data. The goal is to allow the reader to grow comfortable with these tools, encouraging their use and advancing research. We define the research in informatics at the cusp of biotechnological advances in various fields. We introduce the strengths of each technology and encourage its use by describing the opportunities that it offers. We also discuss computational approaches and how they can be applied to various types of data, ranging from the traditional sequence and expression analyses to image-based and proteomics data.

- ***Chapter 5: Image Analysis.*** Image based-analyses are a growing media for biomedical research ranging from basic biology produced in laboratories to patient images obtained in clinical settings. In this chapter, we present computational techniques used to accomplish and automate a plethora of tasks traditionally performed through manual curation. How can I automatically count the number of cells on my plate? Which segmentation algorithm is appropriate for my image?
- ***Chapter 6: Expression Data.*** Microarrays have become commonplace for measuring levels of gene expression in a biological sample of interest. In this chapter, we describe the technology behind microarrays and how it can be used effectively by

a researcher. This chapter will help answer the following questions: How do I normalize and preprocess my data? Which gene or group of genes is significantly up-regulated or down-regulated in my sample? When do I use other technologies like RNA-seq or qRT-PCR to validate my findings?

- ***Chapter 7: A Gentle Introduction to Genome-Wide Association Studies.*** With the advance of genotyping technology, it has become possible to study the effects of genetic variation on various phenotypes of interest. In this chapter, we introduce genome-wide association studies (GWASs) and show how the approach can be used to find novel gene–disease relationships. You will be able to answer the following: Which statistical techniques do I use to detect a genetic association? How can population stratification affect my ability to detect real signal? How do I distinguish between a genetic marker and a causal variant?
- ***Chapter 8: Next-Generation Sequencing Technologies.*** In this chapter, we introduce and describe the broad spectrum of sequencing platforms developed in recent years. We discuss the benefits and trade-offs of various experimental technologies and new computational challenges that have arisen with the advent of next-generation sequencing. How can we use sequencing technologies to perform different types of variant calling? How and when is sequencing used beyond primary sequence analysis (i.e., the transcriptome, ChIP-seq)? Which tools are available for analysis of my sequencing data?
- ***Chapter 9: Proteomics.*** Direct measurement of protein levels provides the researcher with a true snapshot of a biological system. Mass spectrometry and flow cytometry are two common ways of measuring levels of protein expression in a sample. In this chapter, we describe these technologies as well as computational methods used in analysis of such data. How can I measure the abundance of a protein of interest in my sample? What technique can I use to identify subpopulations of cells that alter protein expression in different experimental conditions?

Augmenting Your Data

The final section of this book provides the reader with ways of augmenting their primary data that lead to better scientific understanding. Integrating different data types can provide a deeper insight into biomedical questions. Although these "data mashups" have been presented in a variety of fields, the adoption by biomedical researchers can only be accomplished by providing an understanding of the basic principles used by these methods. The series of chapters that follows shows the reader how to enrich his or her data set with available data repositories. By providing ideas and tools for meta-analysis and data integration, we help the reader identify relevant publicly available resources to further support their findings and generate novel hypotheses.

- ***Chapter 10: Knowledge Base–Driven Pathway Analysis.*** We introduce the notion of biological networks and define a computational framework to model complex

systems by applying Bayesian techniques. This type of modeling can enhance primary analysis by aggregating the observed signals and help identify the underlying biological processes. Which available databases can I use to study my pathway of interest? What computational techniques can I use to build a protein interaction network from my data?

- ***Chapter 11: Learning Biomolecular Pathways from Data.*** Our primary mission here is to educate the reader regarding the general principles and concepts underlying knowledge-base-driven pathway analyses. We consider the challenges presented by high-throughput profiling technologies, for example, used in protein microarrays and metabolomics studies; here we focus on gene expression analyses. How can knowledge-base-driven pathway analysis be used to extract biological meaning from a list of differentially expressed genes and proteins?
- ***Chapter 12: Meta-Analysis and Data Integration of Gene Expression Experiments.*** The research community maintains a rapidly increasing number of repositories containing data from previously published studies. Researchers are able to increase the power of their studies by leveraging these resources through combining several data sources together with their primary data. How can an existing experiment be used to validate my findings without performing any further biological replicates? How can I ask a novel biomedical question by aggregating signal from various data sources?
- ***Chapter 13: Natural Language Processing: Informatics Technique and Resources.*** Knowledge of the past 50 years of biomedical research is captured in an enormous body of literature. Computational approaches have been developed to mine literature sources effectively to enhance and validate findings. This chapter covers the underlying principles of natural language processing used to extract and store knowledge from free text. What controlled terminologies are relevant to my research? Which tools are available for performing natural language processing in the context of my research?

As we have seen, informatics can vastly assist advancement in research and development in biology and medicine. We hope you will find this guide to bioinformatics both accessible and informative regarding existing tools and resources. Furthermore, we believe that this book will build a foundation that allows you to identify and apply additional computational methods to enhance your research. It should exist as a resource allowing you to understand and evaluate cutting-edge bioinformatics approaches encountered in the literature.