Preface

At some time during the course of any bioinformatics project, a researcher must go to a database that houses biological data. Whether it is a local database that records internal data from that lab's experiments or a public database accessed through the Internet, such as NCBI's GenBank, researchers use biological databases for multiple reasons.

One of the original reasons for initiating the fields of bioinformatics and computational biology was the need for biological data analysis. In the past several decades, biological disciplines, including genetics, molecular biology and biochemistry, have generated massive amounts of data that are difficult to organize for efficient search, query and analysis. If we trace the histories of both database development and the development of biochemical databases, we see that the biochemical community was quick to embrace databases. For example, E. F. Codd's seminal paper, "A Relational Model of Data for Large Shared Data Banks" published in 1970 is heralded as the beginning of the relational database, while the first version of the Protein Data Bank (PDB) was established at Brookhaven National Laboratories in 1972.

Since then, and especially after the launching of the human genome sequencing project in 1990 [217], biological databases have proliferated, most embracing the World Wide Web technologies that became available in the 1990s. Now there are thousands of biological databases, with significant research efforts in both the biological and the database communities for managing these data. There are conferences and publications solely dedicated to the topic. For example, Oxford University Press dedicates the first issue of its journal *Nucleic Acids Research* (NAR) every year specifically to biological databases. In 2013, this 20th annual database issue exceeded 1300 pages and included 88 new online molecular biology databases. Among the new databases presented is one containing transcriptome profiling (RNA-seq) data for monkeys, providing new cellular mechanism information about our closest relatives. The NAR database issue is supplemented by an online collection that lists 1,512 databases in 14 categories and 41 subcategories, including both new and updated ones. In addition, Oxford University Press publishes the open access journal, *DATABASE: The Journal of Biological Databases and*

Curation, which contains articles describing novel biological databases and software tools designed to interact with these databases.

Biological database research now encompasses many topics, such as biological data management, curation, quality, integration and mining. Biological databases can be classified in many ways, from the topics they cover, to how heavily annotated they are or which annotation methods they employ, to how highly integrated their databases are with other databases. Popularly, the first two categories of classification are used most frequently. For example, there are archival nucleic acid data repositories (e.g., GenBank) as well as protein sequence motif and domain databases that are derived from primary data sources.

Modern biological databases comprise not only data, but also sophisticated query facilities and bioinformatics data analysis tools; hence the term "bioinformatics database systems" is often used. In this book, we endeavor to explain the world of bioinformatics database systems.

The book is divided into seven chapters. Chapter 1 summarizes the popular and innovative bioinformatics repositories currently available. Discussions include popular primary genetic and protein sequence databases, phylogenetic databases, structure and pathway databases, microarray databases and databases that do not fit into the aforementioned categories, also known as boutique databases.

Chapter 2 discusses the data quality and information integration issues currently involved with managing the bioinformatics databases. With many of the bioinformatics databases looking toward interoperability between the databases, issues concerning data quality are arising. It has been said that the relationship between data quality and data standards is not clearly understood. Do data standards necessarily increase data quality? Depending on the type of data standard and the aspects of data quality considered, the answer to that question is variable. Providing for adequate data quality planning is crucial. Discussions include some of the data quality issues observed in the bioinformatics databases. This chapter presents efforts in the data cleaning field to help with the data quality problems in these databases.

Chapter 3 discusses biological data integration issues and demonstrates how data integration can create new repositories to address needs of the biological communities. The chapter also presents typical data integration architectures employed in current bioinformatics databases. The emergence of high-throughput genomic and transcriptomic datasets from different sources and platforms has enhanced our understanding of how these factors influence complex diseases. It is challenging, however, to explore the relationship between different types of such data sets. Tools available now allow a researcher to build his or her own integrated MySQL database from many popular data sources.

In Chapter 4, biological data searching is explored. Biological data exist in many forms. Macromolecules RNA and DNA are comprised of a handful to thousands of nucleotides. Proteins are formed from amino acids. When new RNA, DNA, or protein is encountered, searching through known elements is the first step in trying to learn about the new element. A search is performed based on the primary sequence, secondary structure or tertiary structure. To learn about evolutionary history of an unknown element, a phylogenetic tree search can be conducted in a database of known phylogenetic trees. If a biological sample is suspected of being diseased, searching through known diseased transcripts helps reveal important treatment intelligence.

Chapter 5 presents the topic of biological data mining. General data mining approaches are discussed, as well as specific data mining methodologies that have been successfully deployed in biological data mining applications. Open biomedical ontologies (OBOs), especially gene ontologies (GOs), are discussed in the context of establishing common vocabularies among large and growing numbers of information stakeholders. Two biological data mining case studies are highlighted that illustrate how data, query and analysis methods are integrated into user-friendly systems. The first case study presents a data mining methodology using a stochastic covariance model to predict the presence of evolutionarily conserved RNA secondary structure elements in a genome. The second case study discusses the genome-wide discovery of coaxial helical stackings. RNA tertiary motifs, such as coaxial helical stackings, are recurring substructures within non-coding RNA (ncRNA), and they play critical roles in a variety of nuclear and non-nuclear cellular functions.

Chapter 6 explores biological network inference. Next-generation sequencing (NGS) technologies generate precise information about DNA and RNA sequences. NGS provides much greater gene expression analytical potential as compared with older microarray technologies. As NGS technologies evolve and mature, a cell's transcriptome will be a reliable and highly detailed representation of gene expression. Using reverse engineering approaches, researchers hope to infer a gene regulatory network (GRN) that will help resolve many unanswered questions about cellular activity. Several current software tools and approaches for biological network inference are presented.

Chapter 7 presents biological data processing approaches using cloud-based technologies. The completion of the Human Genome Project at the turn of the 21st Century [217] and the subsequent advances in high-throughput sequencing of genomic and transcriptomic datasets [378] are increasing demands on the computational infrastructure needed for serious genomic research. As an example, modern high-throughput sequencers, such as the Illumina HiSeq series, produce sequencing reads

from 150 bp in length up to a total of 120 Gbp of data per 27-hour run. Future medical advances depend on our ability to process and analyze large and numerous genomic and transcriptomic datasets. As an alternative to building and maintaining a local infrastructure, cloud computing, in many cases, offers on-demand access to flexible computational resources. There is an ever-growing recognition of the power of cloud computing for large-scale and data-intensive biological applications. We review the use made to date of cloud computing for the processing of biological data, and we discuss challenges presented for biological "big data analytics." The use of cloud computing technologies for bioinformatics research is still in its infancy. However, we anticipate rapid adoption of these technologies as more applications become available and cloud computing costs continue to fall.

Building bioinformatics database systems is a challenging but highly rewarding venture. This book provides enough background information and some state-of-the-art techniques in nascent cross-disciplinary fields. The book is targeted toward researchers and developers of bioinformatics database systems. It can also serve as a supplementary text for a one-semester upper-level undergraduate course or an introductory graduate bioinformatics course. We hope the book will inspire students, instructors, researchers and practitioners who use or build modern bioinformatics database systems.