Preface

Mining biological, biomedical, and health data is an emerging area at the intersection between biology and data mining. There has been an exponential increase of data generation in the last 10 years due to the maturation of sequencing technology and other technologies in bioinformatics and healthcare informatics. It is no coincidence that this revolution has coincided with the completion of the Human Genome Project, but other technological breakthroughs and advances have been critical as well. The bioinformatics fields are in an important stage of new breakthroughs on dealing with dramatically increased information in terms of number, size, and complexity. A new term, “Big Data,” is used to refer to such large, and often diverse, complex, longitudinal, and/or distributed data sets. Big data is usually too large and complex to be processed using traditional data processing applications. It is not so much that the data cannot be processed using traditional tools but that it cannot be processed in reasonable time without new and much faster processing tools. There is and will continue to be much research in this area. As Dr. Shah (2012) stated, “Data-centric approaches that compute on massive amounts of data (often called “Big Data”) to discover patterns and to make clinically relevant predictions will gain adoption. Research that bridges the latest multimodal measurement technologies with large amounts of electronic healthcare data is increasing; and is where new breakthroughs will occur.”

The biological and biomedical research fields are at a revolutionary stage with the use of large, diverse, distributed, and heterogeneous data sets. These “big data” repositories will lead to critical data-intensive decision making and decision support, including biological prediction and clinical decision making and decision support, at a level never before seen. This book can be used as a research reference or handbook offering novel methods for analyzing large-scale data sets. We wish to provide inter-disciplinary research resources in data mining and bioinformatics in this fast-growing area.

In the near future, even higher throughput technologies will only exacerbate the data overload problem and further heighten the need for new techniques and tools that can intelligently and automatically transform the data into useful information and knowledge. One of the central tenets of all information theories is that “the higher the data volume, the lower the information (and knowledge) level.” This can be referred to as the Data Overload, Information Underload problem. This problem exists in most fields, not just in Biology. It has been pointed out by experts in almost all fields that involve voluminous data. The crux of the problem is volume. Data processing tools, which convert voluminous raw data to succinct pieces of information (summaries, relationships, patterns, and other “answers”), are needed which can find (data mine) pertinent, accurate information from the raw data and do it in a reasonable amount of time. So the problem is, as it has always been, scalability of data analytics. Scalability is always cited as one of the main, if not the main, challenge in nearly every major address given by prominent information scientists over the past 50 years. It was the principle motivation for the development of the computer in the first place. Everyone seems to agree on this point.
Scalability comes in at least two varieties, cardinality scalability (too many instances) and dimensionality scalability (too many attributes). The scalability problems can be cast in terms of table terminology as too many rows and too many columns.

An important solution for the curse of cardinality has been to select a representative subset of records (instances or rows), then to analyze or mine that subset. The tacit assumption is that the information (relationships, patterns, summaries, etc.) found in the subset applies to the full data set as well. Whereas, that tacit, statistical assumption can be justified in many cases, it is very difficult to justify in others (e.g., in exception mining). A random subset will almost always miss exceptions, since exceptions are, in some sense, of measure zero. Put another way, if the probability is high that sub-sampling will include an exception, then it may be incorrect to call it an exception in the first place.

An important solution for the curse of dimensionality is also to select a pertinent subset of features (columns or attributes). This process is often referred to as feature selection. For the most part, except for domain knowledge related and analytical dimension reduction, the curse of dimensionality is a fact more than a problem. That is, it is difficult to solve it without at least some loss of information. Thus, the problem is to limit the loss of information while dealing with the problem.

Each chapter in this volume addresses each of these problems in its own way.

Data mining or knowledge discovery in biological databases aims to discover useful patterns from large data volumes. A data mining system is considered linearly row scalable if, when the number of rows is increased by, for example, 10 times, it takes no more than 10 times as long to execute the same data mining process. A data mining system is considered linearly column scalable if the data mining execution time increases linearly with the number of columns or attributes or dimensions.

Data mining, in its most restricted form, can be broken down into three general methodologies for extracting information and knowledge from data. These interrelated methodologies are Association Mining, Classification, and Clustering. Many of the techniques in this book fall in one of these three categories directly, and others attempt to accomplish the same information extraction results in novel and interesting ways.

Association Mining, in short, is a matter of discovering strong association relationships among the subsets of the data. Often these subsets are subsets of columns in the schema. If these associations are unidirectional, the process is called Association Rule Mining or antecedent-consequent relationship mining. If the relationships are undirected, this is called Correlation Mining.

Classification is a matter of discovering signatures for the individual values in a specified column or attribute (called the class label attribute, which can be composite), from values of the other attributes, which are called the feature attributes, in a table, usually called the training table. Classification of numeric tables is usually referred to as prediction. Expert systems that use collections of expert opinions collected over time to evaluate new situations are of this data mining type.

Clustering is a matter of using some notion of instance similarity to group together training table rows so that within a group (a cluster) there is high similarity and across groups there is low similarity. In Biological Data Mining, it is very common to use clustering to accomplish classification (class discovery). That is, when some portion of the data is already classified, the entire data set can be clustered based on some similarity notion. Then unclassified samples can be assigned likely classes based upon the preponderance within its cluster. In Biology this is called putative annotation. The so-called BLAST technologies fall in this category.
For this book, *Big Data Analytics in Bioinformatics and Healthcare*, we have collected cutting-edge research topics and methodologies on managing, analyzing, visualizing, and extracting information from large, diverse, complex, longitudinal, and/or distributed biological and biomedical data sets. Throughout this book you will see, either explicitly or indirectly, the general concepts of association mining, classification or prediction, and clustering. This book, as a collection of research papers, will provide research resources to researchers, practitioners, students, and others in mining massive and complex bioinformatics data sets. This research is important for data-intensive application in the biological and medical fields.

This book encompasses the following three distinct sections: 1) Big Data Analysis Methods and Applications; 2) Reviews and Perspectives on Big Data Analysis; 3) Issues and Concerns in the Big Data Era. The following paragraphs summarize each section and provide an overview for each chapter.

In the first section on Big Data Analysis Methods and Applications, a wide array of new and exciting methods for analyzing big data are presented with special emphasis toward bioinformatics and healthcare. The selected chapters give an, albeit incomplete but well represented, selection of ongoing research questions in the field. First, Yan Guo, Shilin Zhao, Margot Bjoring, and Leng Han open the section with the important topic of RNAseq data analysis. The authors utilized their biological expertise to pinpoint data mining opportunities in RNAseq data analysis, including analysis techniques, tools, and challenges. Then Boya Xie, Qin Ding, and Di Wu discuss an issue faced by many researchers, but especially by researchers in these fields, that of finding the relevant literatures from the various and ever growing databases. These contributors develop a rule-based text mining system, miRCancer, to specifically search for literature in the microRNA profiling area. Although the techniques are specifically focused on searching the literature in microRNA profiling, the novel techniques are very generally applicable and should be of interest to a wide audience in bioinformatics and healthcare data analysis. Next, Zhecheng Zhu, Heng Bee Hoon, and Kiok-Liang Teow discuss the emerging trend of utilizing interactive visualization methods in healthcare systems and the several advantages it has over static visualizations, that is, more flexibility and user-control. A picture is worth a thousand words, it is said. Interactively, moving among pertinent pictures at the user’s discretion must then be even more worthwhile. The chapter focuses on these types of results and information presentation techniques and styles. The chapter contains four case studies where interactive data visualization is applied to various aspects of healthcare systems. This is followed by a report by Anasua Sarkar and Anamika Basu, which focuses on the construction of Gene Networks (GNs) from large-scale gene expression data sets. These authors use clustering and correlation methods to reconstruct GNs from microarray data generated from different stages of embryogenesis in *Arabidopsis thaliana*. Networks are data structures that can hold a wealth of information very succinctly. The results discussed in this chapter should be applicable to a wide variety of other application areas in bioinformatics and healthcare informatics as well. The next chapter of section one is devoted to detection and employment of biological sequence motifs. In this chapter, Marjan Trutschl, Phillip Kilgore, Urska Cvek, Rona Scott, and Christine Birdwell demonstrate a set of tools for detecting, curating, and visualizing sequence and amino acid motifs. Biological sequence motifs are highly conserved across species, thus they are important in many regulatory processes and have structural implications. The next chapter by Paulo Fazendeiro and José Valente de Oliveira presents a new clustering method, fuzzy C-means with a variable focal point, to analyze gene expression data. The FCMFP algorithm is formulated based on the observation that the visual perception of a group of similar objects is dependent on the observer’s position or perspective. The flexibility provided by the fuzzy approach to the well-established C-means
clustering methods provides for customized perspectives depending upon the observer’s preferences. The final chapter of the first section of this book is by Y-H. Taguchi, Akira Okamoto, Yoshiki Murakami, Mitsuo Iwadate, and Hideaki Umeyama. In it, the authors explore a principal component analysis technique based on unsupervised feature extraction methods to identify diseases or biological process associated features in problems, such as detection of aberrant methylation associated with esophageal squamous cell carcinoma, biomarker identification using circulating microRNA, and proteomic analysis of bacterial culturing processes. Principal component analysis has long been known to tease out the important dimensions of raw data, but the application and interpretation is domain specific. This chapter offers specifics on how to do principal component analysis in this specific area.

The middle section of the book is devoted to a very current review of big data analysis from a variety of perspectives. First, Issam El Naqa provides an in-depth review of recent advances in and current challenges to the analysis of big data in radiotherapy. This chapter describes the characteristics of big data in radiotherapy and the future of the emerging field of systems radiobiology for outcomes modeling. This review and perspective is followed by an analysis of genomic data in a cloud computing environment by Philip Groth, Gerhard Reuter, and Sebastian Thieme. They not only discuss how the use of the, so called, cloud can alleviate the burden of storing and analyzing large amount of genomics data but also caution that the challenges, such as infrastructure capability, data protection regulation, risks, and costs, still remains. They make the case that for the ever-increasing pace of data generation in the biology and healthcare sector, cloud computing is becoming an important platform for data analysis in these fields. In a real sense, cloud computing is similar to the much older and time honored “client server” computing model, but it is different as well, in that everything that the cloud provides is larger, more accessible, and more flexible. One could think of cloud computing as providing a complete and unlimited set of services to users who might only have a limited local platform, such as a smart phone or a tablet. All data issues and software needs are provided in the cloud in a time and space independent manner. The next chapter by Ravi Mathur and Alison Motsinger-Reif is more specific. In it, you will find a review of the way in which leveraging biological pathway knowledge can aid in the analysis of large-scale genomic data. This is a prime example of using novel techniques to extract important information from massive raw data stores. Next, Amandeep Kahlon and Ashok Sharma explore computational systems biology perspectives with regard to tuberculosis and in particular the related big data challenges and goals. Once again, though there is a specifically focused perspective in this chapter, tuberculosis, the ideas and solutions discussed have wide application. In the next chapter in section two, Ratna Prabha, Anil Rai, and Dhananjaya Singh give a similar but different account of how “big data” has reshaped their respective research fields by enabling researchers to uncover unprecedented amount of information using new data generation and analysis methods. This is yet another example of how specific techniques developed for specific purposes almost always have wider applicability and usefulness. Data on human complex diseases can come from many sources and be of many complex types. In the next chapter, Kristel Van Steen and Núria Malats discuss integration of other sources of data in addition to domain knowledge to build more predictive models for human complex disease data. The case is made that domain-specific knowledge can offer significant additional insights about data compared to computation alone. This chapter reinforces the very important fact, which is often overlooked by the casual user, that most computation-based systems should be used for decision support, not decision making. It is rare that a computational system is capable
of independent decision making, especially in the bioinformatics and healthcare fields. Associations are relationships that hold much information for disease analysis. In the next chapter, Jami Jackson and Alison Motsinger-Reif review candidate gene and genome-wide genotype and disease associations. Not only is there information in data but there is also information in relationships. Association data mining deals with this fact and is the focus of this chapter. The final chapter in this middle section is devoted to personalized disease phenotypes from massive OMIC data. In it, Hans Binder, Henry Wirth, Lydia Hopp, and Kathrin Lembcke use self-organizing maps to analyze massive molecular medical data and present worked-out examples from different diseases of the colon ranging from inflammation to cancer. This final chapter of the section treats a further step toward the ultimate goal of identifying disease-associated genes, of tailoring disease treatment to the individual. It considers the very important topic of personalized medicine. Self-organizing maps offer both a high degree of information extraction as well as a succinct visualization of the extracted information.

The final section of the book treats some additional bioinformatics and healthcare issues that have emerged in the big data era equally as important as the fundamental data organization and information extraction issues. It addresses topics such as intellectual property and the benefits and concerns of utilizing big data in these sectors. Intellectual property management is critical if the current pace of progress is to continue. There are so many new ideas being developed and in need of IP protection that this is a big data area in its own right. Toward that end, Matthew Knabel, Katherine Doering, and Dennis Fernandez examine controversies and problems associated with intellectual property protection in the area of synthetic biology. Intellectual property protection is critically important in this area, but so is exposure and sharing so that the area can grow and problems can be solved. This trade-off and other issues make the topic very important in today’s world of bioinformatics and healthcare informatics. In the next chapter, Jane Moon, Mary Galea, and Megan Bohensky study clinical data linkages in spinal cord injuries. Even though they are somewhat narrowly focused on data collected in Australia, the issues and results are not. The same issues exist and solutions apply throughout the world. In addition, the solutions can generalize to other types of injuries as well. Finally, in the last chapter of this last section, Andrea Darrel, Margee Hume, Timothy Hardie, and Jeffery Soar discuss the utilization of biological big data by government consortiums in the healthcare sector. These impacts have been overwhelmingly positive. Government involvement has been a very important stimulus and focusing element in this area.

Analysis of big data in the bioinformatics and healthcare sectors has been instrumental in the development of products and solutions for at least the past decade. It has allowed researchers and healthcare providers to have information that was formerly inaccessible. An important trend driving this success is the fact that the cost of data generation continues to decrease. For example, it used to cost $1 billion to sequence a human genome, but currently it costs less than $1000. With such affordable pricing, it is likely that we will soon routinely have our DNA sequences information as part of our medical record. Meanwhile, various hospital-led initiatives have already begun to record patients’ health information in centralized databases as part of patients’ Electronic Medical Records (EMRs). Analysis of big data in these areas is essential for the advancement of the science of healthcare. The diverse and comprehensive coverage of topics on big data analysis in this publication will help researchers, practitioners, and students to better understand the research, methodologies, and discoveries in these all-important fields. We hope this book will be informative to the readers and contribute to the continued growth of big data analysis in bioinformatics and healthcare.
Baoying Wang  
Waynesburg University, USA

Ruowang Li  
Pennsylvania State University, USA

William Perrizo  
North Dakota State University, USA

REFERENCE