Preface

The completion of the sequencing of the human genome marked the beginning of the new era of the so-called *Fourth Paradigm* (Hey, Tansley, & Tolle, 2009) of data-intensive scientific discovery for biomedicine. In addition to the traditional study of living organisms in terms of their biological structure, function, growth, origin, evolution, distribution, and taxonomy (Avila, 1995), biologists have also significantly stepped up their efforts in uncovering the biological processes that underlie disease pathways in the clinical contexts to discover new drugs to benefit human beings. This has resulted in a flood of biological and clinical data ranging from genomic sequences, DNA microarrays, protein structures, and biomolecular interactions, to Gene Ontology, disease pathways, biomedical images, and electronic health records. We are in a scenario where our capability to generate biomedical data has greatly surpassed our ability to mine and analyze the data effectively. In this new era of data-intensive scientific discovery, the focus of biomedical research has shifted from data generation to knowledge discovery, or more specifically, *computational knowledge discovery*. The aim is no longer on enabling the biologists to generate more data rapidly, but on converting them into useful knowledge.

However, in order to translate the vast biomedical data into useful clinical applications, there are still fundamental data analysis difficulties that have to be overcome. Practical issues such as handling noisy and incomplete data (e.g. Protein-Protein Interaction data, or PPIs, having high false positive and false negative rates), processing compute-intensive tasks (e.g. large scale graph mining), and integrating various data sources (e.g. linking genomic and proteomic data with clinical databases) are new challenges faced by biologists in the data-intensive post-genome era.

The objective of this book is to disseminate the research results and best practice from cross-disciplinary researchers and practitioners working on computational knowledge discovery for bioinformatics. By doing so, we hope to bring better awareness of the interesting and challenging problems to inspire new computational solutions. The book is therefore aimed at three overlapping audiences: (1) researchers in the areas of bioinformatics, data mining, machine learning, and data structure; (2) practitioners in the industry in these areas; (3) instructors and postgraduate students in colleges and universities.

Upon reading this book, we hope that researchers in the life sciences will be convinced that *in silico* knowledge discovery tools can truly complement their work *in vitro* and *in vivo*. At the same time, we also hope that computer scientists, mathematicians, and statisticians will be attracted to the new data analysis challenges that are important for computational knowledge discovery in bioinformatics research, and join their life sciences colleagues in their impactful research to benefit humankind.
Computational knowledge discovery is an increasingly important research area in the computer science domain that focuses on the methodologies and processes for automatically or semi-automatically extracting useful patterns or knowledge from large volumes of data. It is often also a cross-disciplinary practice, as the methodologies and processes are applied to discover knowledge in a discipline outside of the computer science domain that has become significantly data-rich—in our case, the life sciences. Its importance has clearly risen with the rapid growth of online information on Web as well as abundant databases used by industries (e.g. various biological or clinical databases), together with the increasingly urgent need to derive novel knowledge from the data generated.

Computational knowledge discovery is itself a challenging task, involving research from multiple computational disciplines such as machine learning, statistics, artificial intelligence, databases, pattern recognition, and data visualization. Generally, it consists of an iterative sequence of the following steps (Han & Kamber, 2005):

1. **data cleaning**: removal of noise and inconsistent data from the original data;
2. **data integration**: combining multiple data sources consistently;
3. **data selection**: identifying and retrieving only the data that are relevant to the analysis task from the database;
4. **data transformation**: transforming and consolidating data into a format that is appropriate for mining by performing summary or aggregation operations;
5. **data mining**: applying intelligent machine learning methods to extract data patterns—this is the key process in computational knowledge discovery;
6. **pattern evaluation**: identifying the truly interesting patterns that represent useful knowledge based on interestingness measures;
7. **knowledge presentation**: using intuitive visualization and effective knowledge representation techniques to present the discovered knowledge to the user.

Steps 1-4 are typically considered as **data pre-processing steps** for preparing the given data for the crucial data-mining task in Step 5 (Fayyad, Gregory & Padhraic, 1996). Steps 6 and 7 are often referred for decision support as they involve choosing interesting patterns/knowledge and presenting them to user in a user-friendly manner for users to do further studies as well as decision-making.

Let us look into more details of Step 5, which is the key process for computational knowledge discovery. Data mining typically involves one or more of the following tasks:

- **Association rule mining (Dependency modeling)**: This task detects relationships between variables/features. One well-known example is market basket analysis in which patterns in supermarket customers’ purchasing habits are extracted from their sales transactions (“market baskets”) and translated for market campaigns. For example, the association rule \{butter, bread\} => \{milk\} can be detected in most supermarket transaction data—it indicates that if a supermarket customer has bought butter and bread together, he or she is likely to buy milk as well. With such association rule mining, a supermarket can automatically detect which products are frequently bought together and use this information for marketing purposes, e.g., promotional pricing or product placements. In the biomedical domain, association rule analysis can be performed on gene expres-
sion values to mine for association rules where the antecedents could be features and their value ranges (cancer genes and corresponding gene expression values under different conditions) and the consequents the class labels (cancer or non cancer). The knowledge discovered can then be used for building a diagnostic system.

- **Cluster analysis or clustering**: This is the task of segmenting a set of objects into groups so that the objects in the same group are more similar to each other than to those in other groups. In machine learning, cluster analysis is considered a form of *unsupervised learning*, since there is no need for users of the clustering algorithms to provide training examples. One example use of cluster analysis in biomedical research is to cluster gene expression data into groups where genes in each group share similar gene expression profiles, to discover genes that belong to the same biological functions (Li, Tan & Ng, 2006).

- **Classification**: This is the task of assigning input data into one of a given number of categories. A classic example is spam filtering, in which the task is for an email program to classify a new email as a legitimate message or just spam. To build a classifier, a user must first collect a set of training examples that are labeled with the predefined known classes (for example, known spam messages and different types of legitimate email messages). A machine learning algorithm is then applied to the training data to build a classification model (classifier) that can be employed subsequently to assign the predefined classes to instances in a test set (for evaluation) or future instances (in practice) (Li, Liu, & Ng, 2007). An application of classification in biomedical research is to predict the biological functions of novel proteins or genes. Here, those proteins with known functions are used as training data to build a classification model for each biological function, which can be subsequently used to classify unknown proteins.

- **Regression analysis**: This task attempts to find a function, which models the data with the least error. It includes the techniques for modeling and analyzing several variables, where the focus is on the relationship between a dependent variable and one or more independent variables.

- **Anomaly detection (Outlier detection)**: In this task, we attempt to detect data records, which do not conform to an expected or established normal behaviour. The mined results might be either interesting data records or error records, which require further investigation.

**COMPUTATIONAL CHALLENGES**

Let us now discuss in details the various data analysis challenges for computational knowledge discovery from biological data (Li & Ng, 2009).

**Handling Noisy and Incomplete Data**

Biologists have generated tremendous amounts of biological data with the newly available advanced high-throughput technologies, but many of the data sets are actually rather noisy and incomplete. For example, there is a high-level of noises in gene expression data detected by DNA microarrays and protein interactions detected by popular high-throughput assays such as yeast-two-hybrids. In addition, molecular structures (e.g. protein structures), molecular interactions (e.g. protein interactions), Gene Ontology (including the sub-ontologies, molecular function, biological process, and cellular component), and disease pathways are still largely incomplete due to relatively limited biological knowledge.
Let us take Protein-Protein Interactions (PPIs) as an example to illustrate that current high-throughput data sets can be very noisy (alarmingly high false positives) and incomplete (alarmingly high false negatives). Established high-throughput methods such as Yeast-Two-Hybrid (Y2H) (Fields & Song, 1989) and Tandem Affinity Purification-Mass Spectrometry (TAP-MS) (Puig, et al., 2001; Rigaut, et al., 1999) are popular methods that have enabled comprehensive detection of PPI data sets in various species. However, the quality of detected PPI data sets is far from satisfactory.

- **Noisy data:** The experimental conditions in which the above detection methods are carried out may result in detecting interactions that do not occur under *in vivo* physiological conditions. In other words, the experimental data may not be very accurate and contain noisy false positive interaction data that do not occur in the cell.

- **Incomplete data:** The high-throughput methods can also fail to detect various types of interactions. For example, researchers have reported loss of weak transient interactions, loss of post-translational modification, and bias against soluble or membrane proteins (Lalonde, et al., 2008; Tarassov, et al., 2008). These result in false negatives in the PPI data leading to incomplete coverage of the interactomes.

How serious is the data quality situation for protein interaction detection—specifically, what are the false positive and false negative data rates in current PPI data sets? Many researchers have attempted to answer this question (Bader & Hogue, 2002; Bader, Chaudhuri, Rothberg, & Chant, 2004; Gentleman & Huber, 2007; Hart, Ramani, & Marcotte, 2006; Von Mering, et al., 2002) by comparing the overlaps between the collected protein interactions from different large-scale biological experiments, or checking the consistency of the function or location information between the interaction partners, to estimate the accuracy and coverage of the PPI data. The results of all these investigations have consistently shown that the quality of the protein interaction data is indeed very problematic, with accuracy ranging from 10% to 50% (Gentleman & Huber, 2007; Sprinzak, Sattath, & Margalit, 2003; Von Mering, et al., 2002) and coverage lower than 50%, even for the most studied and curated interactome of *Saccharomyces cerevisiae* (yeast) (Hart, et al., 2006). One can only expect the accuracy and coverage of the PPIs for the other less-studied or more complex species to be even lower.

Can we employ computational knowledge discovery methodologies to help dealing with the experimental limitations of noisy and incomplete protein interaction data? To address the issue of noisy data or false positive interactions, we can computationally validate the experimental protein interactions using various robust reliability measures (Chen, Hsu, Lee, & Ng, 2006). To address the issue of incomplete interactome coverage or false negative interactions, we can computationally predict novel protein interactions to supplement the experimental data set of detected PPIs. One way to do this is to represent the known protein interactions (positive training examples) and the non-protein interactions (negative training examples) using various informative biological features/properties, such as protein domains, protein sequences (e.g., amino acids composition, etc.), protein functions, biological processes, cellular locations, structural information, and topological features extracted from the PPI network. Then, a supervised machine-learning model (such as support vector machines, Bayesian classifier, k-NN classifier, or decision tree) can be used to classify unknown protein pairs into interacting or non-interacting. Another approach is to develop network-cleansing techniques such as finding missing links and/or removing unreliable links, which can also be used to improve the quality of the entire protein interaction networks (Li & Ng, 2009).
A number of chapters in this book address the noisy and incomplete data issues in protein-protein interaction detection (Section 3). To address the incomplete data issue, Chapter 10, “Prediction of Protein-Protein Interactions between Human Host and Two Mycobacterial Organisms,” predicts host-pathogen interactions to help biologists better understand disease processes. Chapter 12, “Incorporating Network Topology Improves Prediction of Protein Interaction Networks from Transcriptomic,” introduces a novel approach Function Restricted Value Neighborhood (FRV-N) for the identification of potential protein-protein interactions based on the integration of known PPI network topology and transcriptomic data.

To address the noisy data issue, Chapter 11, “A Transfer Learning Approach and Selective Integration of Multiple Types of Assays for Biological Network Inference,” proposes a transfer learning algorithm to integrate noisy interaction data to infer more reliable networks. Chapter 13, “Mining Protein Interactome Networks to Measure Interaction Reliability and Select Hub Proteins,” describes a k-round signal flow simulation algorithm to measure interaction reliability from connection patterns of the interactome networks.

**Processing Computer-Intensive Network Mining Tasks**

In order to uncover the biological processes of how life functions in living organisms, it is important to investigate the biological interactions at the network level. Large-scale biological networks include gene regulatory networks, metabolic networks, protein-protein interaction networks, protein domain networks, as well as DNA-protein interaction networks. Given that proteins are the workhorses of biological processes and they function only through interactions, the most intensively studied networks are naturally the PPI networks.

Computationally, a PPI network is typically modelled as an undirected graph (sometimes with weighted links) where the nodes represent unique proteins and the links denote interactions between two proteins. Such a network is a very large graph with thousands of vertices and tens of thousands of edges even for a simple model organism such as yeast. One can only imagine the insurmountable complexity of the PPI networks for the more complicated species such as the human being.

Graph theory is an important tool to facilitate efficient analyses of the large-scale interaction networks (Barabasi & Oltvai, 2004). Some important graph mining techniques have been proposed and applied in discovering useful biological knowledge from PPI networks such as interaction motifs, network motifs (Alon, 2007), lethal proteins (Jeong, Mason, Barabsi, & Oltvai, 2001), protein complexes/functional modules (Li, Wu, Kwoh, & Ng, 2010; Li, Tan, Foo, & Ng, 2005; Xie, Kwoh, Li, & Wu, 2011), and disease genes (Yang, Li, Wu, & Kwoh, 2011). We can also investigate the evolution of various protein interaction networks by computing the alignment of the PPI networks. The computational challenge for such a task is clearly humongous, as we will need to handle multiple massive networks. Once the networks are aligned, we can transfer knowledge (e.g. protein interactions, functions, etc.) from a PPI network for well-studied species to other less well-studied species.

Clearly, mining the large protein interaction networks is a compute-intensive task. Techniques for reducing the search space and time complexity are therefore highly important, and we can obtain better graph mining results by exploiting biological knowledge coupled with the development of novel efficient graph mining techniques.

The chapters in Section 4 cover various topics in biological network mining. Chapter 15, “New Trends in Graph Mining: Structural and Node-Colored Network Motifs,” provides a technical overview

Chapter 14, “Finding Minimum Reaction Cuts of Metabolic Networks under a Boolean Model Using Integer Programming and Feedback Vertex Sets,” introduces an integer programming-based method to find an optimal set of reactions to be inactivated in a large biological network, such as the E. coli metabolic network. The proposed method can find an optimal set of reactions to be inactivated much faster than a naive IP-based method and several times faster than a flux balance-based method.

Chapter 16, “Discriminative Subgraph Mining for Protein Classification,” studies the graph classification problem for protein classification. Unlike existing methods, which use frequent subgraphs as features (which results in a large search space), this chapter presents two efficient discriminative subgraph mining algorithms: COM and GAIA, which directly search for discriminative subgraph patterns instead. Experimental results showed that COM and GAIA can achieve high classification accuracy and runtime efficiency.

**Integrating Various Biological Evidences for System-Level Understanding**

As mentioned, the recent advent in high-throughput experimental technologies has resulted in an increasing number of large datasets available at various biological levels from the genome to the proteome and metabolome. Many of the datasets are deposited in centralized databases that are publicly accessible by all the researchers. At the same time, there are also increasing efforts by biologists to provide annotations of biological knowledge, with projects such as the Gene Ontology initiated to enable systematic functional annotations of genes and proteins. These efforts were motivated by the need to integrate various biological evidences in order to achieve system-level understanding of the complex cellular processes.

Let us continue to use protein interactions as examples. There are at least two key benefits of integrating various biological evidences. Firstly, by integrating multiple data sources together, we can derive better quality data. As mentioned, the PPI data generated from the high-throughput methods are inevitably noisy and incomplete. It will be useful if we can weigh the links (i.e. the protein interactions) in the PPI networks by using appropriate confidence measures. For example, we can employ metrics such as reproducibility of the interactions from multiple experimental methods, support from such other non-interaction data as co-expression, co-localization and shared functions, as well as the conservation of the protein interactions across other genomes, etc. There are various ways to integrate biological data together. For example, in one of the works to predict protein interactions (Jansen, et al., 2003), a Bayesian network model was developed by integrating noisy experimental interaction data with the weighted genomic features that are only weakly associated with interactions, such as messenger RNA co-expression, co-essentiality, and co-localization. Another way is to use machine learning methods, such as kernel methods (Ben-Hur & Noble, 2005), to integrate different biological resources into high dimensional vector space to do classification for better prediction of protein interactions.

Secondly, integrating the PPI networks with additional biological evidences will exploit the rich information and strengths from the different sources at different biological levels (Ghazalpour, Doss, Zhang, Wang, & Plaisier, 2006; Kelley & Ideker, 2005; Mootha, et al., 2003) to provide a system-level understanding of biological processes and human diseases (Ideker & Sharan, 2008). For example, successful development of diagnostics and therapeutics that target disease-relevant protein interactions can only be attained by obtaining quantitative and dynamic PPI data across different tissue cells and
integrating them with gene expression, functional, structural, and metabolic pathway data to provide a system-level picture.

Nevertheless, one should be mindful that data integration may not always result in high-quality data and results—the intrinsic noisiness of data and integration methods do matter. Computational knowledge discovery methods based on machine learning approaches, probabilistic, and statistical tools should be carefully designed to maximally exploit the knowledge from various biological evidences, while minimizing the side effects coming from the irrelevant and noisy sources.

There are two chapters in this book that cover the topic of integrating multiple data sources for system-level understanding. Chapter 5, “Clustering Genes Using Heterogeneous Data Sources,” elaborates on how to integrate the incomplete data sets (e.g. from the literature) to enhance clustering algorithms. Chapter 12, as mentioned earlier, integrated known PPI network topology and transcriptomic data from 170 yeast microarray profiles (Affymetrix ‘GeneChips’) for identifying potential protein interactions.

ORGANIZATION OF THE BOOK

This book is organized into five major sections, covering the following topics: Sequence and Function analysis (Section 1), Gene Expression Data Analysis (Section 2), Protein Interaction Analysis (Section 3), Network Mining and Systems Biology (Section 4), and Biological Data Mining for Healthcare (Section 5).

Section 1: Sequence and Function Analysis

Motif identification for DNA sequences (including single group and two group identification) has many important applications in biological studies. In Chapter 1, “Identification of Distinguishing Motifs,” the authors introduce a single-group algorithm that allows indels (insertions and deletions) in the occurrences of the motif to accommodate the potential errors in the occurrences of the motif (current methods only focus on the single group and do not allow indels). An algorithm for the two-groups problem is also given along with extensive simulations evaluating algorithms.

One limitation faced by biologists when designing microarray experiments is that current probe mapping techniques only allow one mismatch between probe and genomic sequences. Chapter 2, “Mapping Affymetrix Microarray Probes to the Rat Genome via a Persistent Index,” describes how a novel database implementation of a $q$-gram index can address this practical issue.

Chapter 3, “Improving Prediction Accuracy via Subspace Modeling in a Statistical Geometry Based Computational Protein Mutagenesis,” presents a computational approach for representing a system of protein mutants, due to single residue replacements, that are experimentally determined to have either “unaffected” or “affected” activity levels relative to the native protein.

Chapter 4, “Characterization and Classification of Local Protein Surfaces Using Self-Organizing Map,” describes the prediction of proteins’ functions from their tertiary structures on the molecular level. The local surface patches of proteins, which are characterized by their shape and electrostatic potential, are clustered using an emergent self-organizing map approach, resulting in the discovery of 30-50 clusters of local surfaces of different characteristics.
Section 2: Gene Expression Data Analysis

Standard exploratory clustering algorithms have been used to cluster gene expression data for identifying closely related genes. The algorithms typically assumed that multiple and complete sources of data are available. In Chapter 5, “Clustering Genes Using Heterogeneous Data Sources,” the authors present a new clustering algorithm to deal with multiple complete and incomplete information sources. This is important, as many incomplete sources of data are also likely to be useful in their inclusion for gene expression data analysis.

Chapter 6, “Efficient Mining Frequent Closed Discriminative Biclusters by Sample-Growth: The FDCluster Approach,” proposes an algorithm FDCluster to mine frequent closed discriminative bicluster in multiple microarray datasets, which can reduce the noise influence and allow more biological biclusters to be found. The experimental results showed that FDCluster is more effectiveness than traditional methods in either a single microarray dataset or multiple microarray datasets. The authors also tested the biological significance using GO to show that their proposed method is able to produce biologically relevant biclusters.

Chapter 7, “Wave-SOM: A Novel Wavelet-Based Clustering Algorithm for Analysis of Gene Expression Patterns,” presents a wavelet-based algorithm (Wave-SOM) to help visualize and cluster oscillatory time-series data in two-dimensional gene expression microarrays. Using previously studied expression patterns of yeast cell cycle and functional genes as test data sets, the proposed algorithm outperformed five other competing programs.

Regulatory network analysis and other bioinformatics tasks require the ability to induce and represent arbitrary Boolean expressions from data sources. Chapter 8, “Mining Frequent Boolean Expressions: Application to Gene Expression and Regulatory Modeling,” introduces a novel framework called BLOSOM for mining (frequent) Boolean expressions over binary-valued datasets. Application studies on gene expression and gene regulation patterns showcase the effectiveness of this approach.

Chapter 9, “Bioinformatics Methods for Studying MicroRNA and ARE-Mediated Regulation of Post-Transcriptional Gene Expression,” provides a review of several miRNA target prediction tools and data sources, as well as the computational methods used for the prediction of AREs. The authors discuss the connection between miRNA and ARE-mediated post-transcriptional gene regulation. Finally, a data mining method for identifying the co-occurrences of miRNA target sites in ARE containing genes is presented.

Section 3: Protein Interaction Analysis

Chapter 10, “Prediction of Protein-Protein Interactions between Human Host and Two Mycobacterial Organisms,” describes a computational method for predicting host-pathogen interactions to help biologists better understand disease processes. Given the recent re-emergence of pathogens such as *M. tuberculosi*, this work is certainly very timely and important to help defend the public against these infectious diseases.

As mentioned, it is well known that biological networks are noisy due to the limitations of biological experiments. Chapter 11, “A Transfer Learning Approach and Selective Integration of Multiple Types of Assays for Biological Network Inference,” demonstrates how a transfer-learning algorithm can be used to integrate noisy interaction data obtained from multiple types of assays to result in more reliable networks.

Chapter 12, “Incorporating Network Topology Improves Prediction of Protein Interaction Networks from Transcriptomic,” introduces a novel approach called Function Restricted Value Neighborhood
(FRV-N) for the identification of potential protein-protein interactions based on the integration of known PPI network topology and transcriptomic data. The proposed method was used to reconstruct PPI networks using an experimental data set consisting of 170 yeast microarray profiles. The positive results demonstrate that incorporating the interactome's topological knowledge can improve the ability of transcriptome analysis in reconstructing interaction networks with a high degree of biological relevance.

Chapter 13, “Mining Protein Interactome Networks to Measure Interaction Reliability and Select Hub Proteins,” describes a k-round signal flow simulation algorithm to measure interaction reliability from the connection patterns of the interactome networks. In addition, the chapter also presents an algorithm for mining the complex interactome network structure by restructuring the network using hierarchical ordering of nodes. The structure re-formatting process is able to reveal hub proteins in the interactome networks, which are key players in the biological processes.

Section 4: Network Mining and Systems Biology

Chapter 14, “Finding Minimum Reaction Cuts of Metabolic Networks under a Boolean Model Using Integer Programming and Feedback Vertex Sets,” introduces an integer programming-based method to find an optimal set of reactions to be inactivated in a large biological network, such as the *E. coli* metabolic network. One potential translational bioinformatics application for this work is the measurement of the structural robustness of metabolic networks for detecting better drug targets.


Protein classification can be performed by representing 3-D protein structures by graphs and then classifying the corresponding graphs. One effective way to classify such graphs is to use frequent subgraph patterns as features; however, the effectiveness of using subgraph patterns in graph classification is often hampered by the large search space of subgraph patterns. Chapter 16, “Discriminative Subgraph Mining for Protein Classification,” presents two efficient discriminative subgraph mining algorithms: COM and GAIA. These algorithms directly search for discriminative subgraph patterns instead of frequent subgraph patterns, which can be used to generate classification rules. Experimental results show that COM and GAIA can achieve high classification accuracy and runtime efficiency.

Incongruence between species trees and gene trees has remained a challenge in molecular phylogenetics for its biological and algorithmic complexities. Chapter 17, “Infer Species Phylogenies Using Self-Organizing Maps,” presents a Self-Organizing Map (SOM) based phylogeny inference method and demonstrates its superiority to the state-of-the-art gene concatenation method using the same datasets. This chapter illustrates how to cluster multispecies genes, estimate multispecies gene entropy, and visualize the species patterns through the self-organizing map mining.

Section 5: Biological Data Mining for Healthcare

Chapter 18, “SPCC_{TDM}: A Catalogue for Analysis of Therapeutic Drug Monitoring Related Contents in the Drug Prescription Information,” describes a text mining method to analyse Therapeutic Drug Monitoring (TDM) in prescription information of drugs (Summary of Products Characteristics, SPC), which consists of six structure-related items (dose, adverse drug reactions, drug interactions, overdose,
pregnancy/breast feeding, and pharmacokinetics), four theory-guided items according to the information about ranges of plasma concentrations, and a recommendation of TDM in the SPC.

Chapter 19, “Hierarchical Density-Based Clustering of White Matter Tracts in the Human Brain,” introduces a new framework for automatic white matter tract clustering using a hierarchical density-based approach. The chapter describes experiments on real data that demonstrate the proposed method can be used to effectively group these tracts into meaningful bundles on multiple scales as well as eliminating noisy fibers.

Chapter 20, “Revealing the Origin and Nature of Drug Resistance of Dynamic Tumour System,” identifies the strategies that resistant subpopulations of cancer cells undertake to overcome the effect of the anticancer drug Topotecan. For the analyses of cell lineage data encoded from time-lapse microscopy, data mining tools that generate interpretable models of the data and provide the necessary statistical significance are chosen.

CONCLUSION

As computing becomes increasingly integrated into the entire pipeline of biological and medical discovery process, there is an increasingly pressing need for computer scientists, mathematicians, statisticians, and bioinformaticians to collaborate and contribute to the development of the life sciences by creating novel techniques discovering useful knowledge from large-scale real-world biomedical data. The opportunities for computational researchers to cross over from the computation domain into the biomedical domain to contribute to the meaningful scientific pursuit with the biologists and clinical scientists are abundant, as shown by the various chapters in this book.

The ultimate success of the computational knowledge discovery for bioinformatics analysis will depend on the parallel improvements both in the biological and clinical experimental techniques from the biologists and clinicians to provide rich and clean biological/clinical datasets for the knowledge discovery community, and in the computational techniques from the computer scientists, mathematicians, and statisticians to provide efficient and effective ways to exploit the data for knowledge discovery. The ultimate goal is to enable biologists and clinicians to better understand life processes. We hope this book will play a role in helping to realize what a cross-disciplinary marriage of biological science with computer science can bring.

Finally, we wish to mention that all the chapters in this book have been selected from Volume 1 of International Journal of Knowledge Discovery in Bioinformatics (IJKDB). As such, our heartfelt thanks go to the members of IJKDB international advisory board, the associate editors, and the members of the editorial review board in reviewing of the manuscripts and providing useful feedback. Of course, we are grateful to the authors who contributed to the exciting and important research topics of developing computational knowledge discovery approaches for applications in biology and healthcare.

Special thanks also go to the publishing team at IGI Global, for providing invaluable contributions and guidance throughout the whole process from inception of the initial idea to the final publication of this book.

Xiao-Li Li
Institute for Infocomm Research, A*STAR, Singapore

See-Kiong Ng
Institute for Infocomm Research, A*STAR, Singapore
REFERENCES


