Knowledge-Based Bioinformatics – From analysis to interpretation.

‘Knowledge-Based Bioinformatics – From analysis to interpretation’, edited by Gil Alterovitz and Marco Ramoni, is a collection (http://onlinelibrary.wiley.com/book/10.1002/9780470669716) of works by 35 authors arranged in 13 Chapters of 4 sections and 2 parts. This book includes diverse knowledge-based bioinformatics issues that arise from rapidly increasing data generated by progressive biological research and high-throughput experimental technologies. In fact, the development and practical use of biological knowledge base is a main issue in bioinformatics literature. This book covers a broad range of topics related to the types of knowledge required for bioinformatics and extensive examples of knowledge databases for understanding of knowledge-based bioinformatics, and knowledge-driven discovery and various data analysis methods for its applications, focusing on representation, interpretation, acquisition, development, integration, maintenance, relationships and networks of biological knowledge.

The book is organized into the following four sections: 'Knowledge-Driven Approaches' (Section 1), 'Data-Analysis Approaches' (Section 2), 'Gene and Protein information' (Section 3) and 'Biomolecular Relationships and Meta-relationships' (Section 4). Sections 1 and 2 introduce ‘Fundamentals’ (Part I), and using this background information ‘Applications’ (Part II) are discussed in Sections 3 and 4. Each section consists of three or four chapters under the same specific topic. At the end of the book, trends on current bioinformatics and conclusions are briefly mentioned.

Chapter 1 begins by providing a review of historical and present knowledge discovery in bioinformatics, ranging from definition for knowledge, to formal reasoning, to knowledge representations (KR). The issues on common knowledge and the capture of novel knowledge are also mentioned. Using this initial background, key topics of knowledge discovery and data mining (KDD) and its application domains are presented from various perspectives such as ontology, text information extraction, gene expression analysis, pathway structure, relation mappings between genotypes and phenotypes, web’s role in knowledge mining, information aggregation and articulation for linked knowledge, and new requirements for the next-generation KDD.

Chapter 2 mentions that the biological researches have moved from high-quality experimental data generation to data analysis, and molecules of interest have moved from single genes and their behaviors or functions to groups of interacting entities. According to this shift, the attendant challenges of knowledge-driven approaches in finding and interacting with biomedical knowledge, e.g. automatically parsing specific meaning and key terms from unstructured text, handling of implicit knowledge and visualizing highly complex data, are addressed. The chapter also describes current knowledge-based bioinformatics tools: (i) enrichment tools such as GO (Gene Ontology), DAVID (Database for Annotation, Visualization and Integrated Discovery) and GSEA (Gene Set Enrichment Analysis) databases for interpreting groups of genes, (ii) network/interaction tools such as DIP (Database of Interacting Proteins) and (iii) much richer network tools such as STRING (Search Tool for the Retrieval of Interacting Genes/Protein) databases for supporting indirect interactions and comprehensive investigation of a group of interesting proteins and their associations with other biological entities. In addition, 3R knowledge-based systems built on three broad classes of methods: reading, reasoning and reporting and the Hanalyzer 3R system architecture are stated.

Chapter 3 deals with ontology, one of major biological knowledge that is a computational formalization to provide consistent descriptions and reasoning schemes for entities in a specific domain and their relations by using controlled vocabularies or precise semantics. The authors points out that ontology building is in a transition stage from a craft to a fully industrial engineering discipline, and current bio-ontologies often reveal limitations in supporting automated reasoning. This chapter examines KR languages for building bio-ontologies such as the Resource Description Framework (RDF), the Web Ontology Language (OWL) and the Open Biomedical Ontologies (OBO) format, and summarizes features of each KR language. Guidelines on best practices for building bio-ontologies are suggested in.
terms of the scope of ontology, identification schemes for biological entities and design patterns.

Chapter 4 discusses the following issues for design/implementation/updating of knowledge bases on the basis of two databases, InterPro and UniProtKB: how data feed into the biological knowledge repositories, how knowledge databases are maintained, what principle factors are taken into account in choosing database architecture, how instances are arranged in the databases, what is the prevalent programming language used in software development project, which interfaces to access data in knowledge bases can be presented to end users, how data in databases are updated, how a multi-source database is handled and how various information are integrated from multiple and disparate servers.

Chapters 5 to 7 mainly explores data analysis. Chapter 5 discusses statistical analysis and learning methods used in bioinformatics. Genomic researchers often confront multiple comparisons of genes or loci across a set of samples. This chapter presents the concepts of significance testing, multiple testing, family-wise error rate (FWER), false discovery rate (FDR) to uncover unexpected patterns or relationships in data. Several high-dimensional clustering methods (k-means clustering and hierarchical clustering) which can be applied to discover groups of genes or other entities with similar expression patterns, and similarity metrics for the clustering (Euclidean, Manhattan and correlation distances) are introduced. It also comprises the data reduction methods [e.g. principal component analysis (PCA), independent component analysis (ICA) and multidimensional scaling (MDS)] to encapsulate the values of the measured variables across all samples with a small number of synthetic variables. For example, the methods can be applied to find principal components to distinguish a specific cancer cell lines from the others. Additionally, classification and prediction algorithms based on machine learning methods [e.g. quadratic discriminant analysis (QDA), naive Bayes and support vector machine (SVM)], which can be applied to predict clinical outcomes from genomic measures are addressed.

Chapter 6 deals with Bayesian methods that have been widely adopted in bioinformatics since they can provide an intuitive way to combine prior information and different types of data in a consistent framework. This chapter introduces Bayes theorem and three key elements in applying the theorem to a specific problem: (i) model specifications to evaluate the probability distribution of the observations conditional on a specific parameter, (ii) prior specifications for the model parameter distribution and (iii) computational methods for observation probability. This chapter also briefly describes statistical inference based on the Gibbs sampling where the posterior distributions are inferred by iteratively sampling from a set of conditional distributions. It also exemplifies some Bayesian inferences of population structure from the collected genetic marker data, for example, inference of protein binding motifs from sequence data, inference of transcriptional regulatory networks from joint analysis of protein–DNA binding data and gene expression data and inference of protein and domain interactions from yeast two-hybrid data.

Chapter 7 covers biomedical text mining based on natural language processing (NLP) techniques. The authors discuss the information type of text-mining systems to assist users, and provide an overview on components of NLP solutions used for text mining (e.g. tokenization, part-of-speech tagging, named entity recognition, parsing, abbreviation resolution, text retrieval, passage retrieval, interaction information extraction and question answering). Resources needed for text analysis, system evaluation metrics and several methodologies applied in practical systems such as text retrieval, term recognition and relationship identification are included with development phases of a text-mining systems and their specific tasks in this chapter. Some selected examples of successive text-mining systems are introduced.

Section 3, ‘Gene and Protein Information’ comprises Chapter 8 through Chapter 10. Chapter 8 explains the fundamentals of gene ontology functional annotation based on the GO database. In this chapter, the following contents related to the GO resource are reviewed: GO terms provided as three separate vocabularies (molecular function term, biological process term and cellular component term), its annotation process (e.g. evidence codes and annotation qualifiers for its manual annotation process, computational annotation and community involvement in the annotation process), its usage, its limitation, GO browsers, GO functional analysis tools (FatiGO, DAVID, Onto-Express and the Ontologizer), GO slims and GO displays.

Chapter 9 describes the genome annotation that intends gene prediction and description of any element on the genome to which a biological
functionality can be attached. In general, the value of a genome sequence depends on the quality of annotation. This chapter starts with the introduction of various strategies for gene finding process and popular gene prediction programs (GENSCAN, NSCAN, GENEID, AUGUSTUS and Ensembl genebuilds), and manually annotated gene sets (HAVANA genes and GENCODE data sets). The topics on comparison between the automated annotation process and the manual approach are explained along with CCDS (consensus-coding sequence) project and annotations for pseudogenes/non-coding RNA genes. In addition, the following problems related to the impact of next-generation sequencing (NGS) on genome annotation are discussed: alignment of sequences between multispecies genomes, web-based community annotations, tag-profiling methods [Serial Analysis of Gene Expression (SAGE), Cap Analysis of Gene Expression (CAGE), RNA-Seq], SNP [single nucleotide polymorphism] or CNV (copy number variation)-based human genome sequence variation, and annotation of unstable gene families.

Chapter 10 focuses on the BL-SOM (batch-learning self-organizing map) method, which is an unsupervised neural network algorithm and can be adapted to classification of genomic sequence fragments on a species basis without orthologous sequence set or sequence alignment. This chapter explains the concept of the BL-SOM learning for classification and its weight vector updating rule, and exemplifies genomic sequence analyses using BL-SOM with the following applications: (i) species-specific separation of eukaryotic sequences, (ii) phylogenetic separation of prokaryotic sequences obtained by metagenome analyses, (iii) phylogeny estimation of genomic sequence fragments of novel microorganisms and (iv) reassociation of fragmental sequences on a species basis.

Section 4 comprises Chapters 11 through 13, focusing on biomolecular relationships and metarelationships. Chapter 11 reviews types of commonly conducted molecular network analysis and their applications, and their challenges/future directions. Specifically, the following four types of network analysis and their related subtopics are presented: (i) topology analysis, network statistics (degree, clustering coefficient, shortest path length, eccentricity and node/edge betweenness centrality), and applications of topology analysis (identification of important nodes in networks or pathways, and network-based protein function prediction), (ii) motif analysis, its concept and methods for motif identification/analysis, (iii) network modular analysis, modular structure identification by using various network clustering methods (density-, partition-, centrality-based and hierarchical clustering), and the applications (protein function prediction and biomarker discovery) and (iv) molecular network comparison, graph matching algorithm (PathBLAST, Ogata’s linear path algorithm, PageRank-like algorithms: IsoRank, IsoRank-Nibble), and applications of molecular network comparison (network alignment, network integration and network querying). Chapter 11 also presents summaries and comparisons of extensive network analysis software and tools commonly used.

Chapter 12 deals with biological pathway knowledge bases and integration of pathway-related annotations. Some of the major issues facing bioinformatics are integration of new research results with previous knowledge about specific biological pathways and visualization to represent the increasing complexity of pathway information. In this chapter, various knowledge bases (Reactome, KEGG, WikiPathways, NCI-PID, NCBI BioSystems, Science Signaling and PharmGKB) for metabolic, regulatory, signaling, disease and drug pathways are compared. The overviews of pathway curation for representing a set of related biological events or reactions in a given biological context are introduced. The data exchange forms for integrating data of different formats (SBML, BioPAX, PSI MI and KGML-ED) without rewriting data and interactive visualization tools are explained. A description of analysis scenario in the Reactome tool and challenges/future directions of pathway knowledge bases are also discussed.

Chapter 13 explains methods and challenges of identifying biomolecular relationships and networks associated with complex traits, and their application to drug treatment. In practice, the understanding of complex traits is crucial to comprehend the status of health and disease. Thus, the inference/estimation process for relationships between genes and phenotypes in complex traits is important to solve the molecular secrets of complex traits. In the chapter, genetic phenomena causing difficulties in identifying and analyzing intracellular networks associated with complex traits, and representation models for relationships between traits and genes are investigated. In addition, extensive reviews on the resources available for integrative systems biology to aid in disease exploration and drug discovery are comprised with
their web link information. The methods combining functional genomics results and existing biological information to build novel biological networks, advantages of networks exploration in molecular biology and drug discovery, and practical examples of network exploration are presented.

Interpreting available knowledge and data generated by high-throughput experimental technologies and finding their molecular interactions for methodical functional analysis are essential roles in current bioinformatics. This book extensively covers such topics in terms of biological knowledge from fundamental to applications by exploring the following three major issues: (i) various kinds of biological knowledge databases, (ii) knowledge-driven and data analysis approaches for curation, acquisition, maintenance, integration, interpretation, relationships and networks of biomedical data and (iii) applications. Thus, curators, industry professionals, students and researchers in various areas such as computer science, mathematics, statistics and biological sciences in the biomedical field will benefit from this book. In particular, this book can be used as an introductory textbook about bioinformatics that readers can learn relevant background information, types of required knowledge, research issues and recent developments about the subject of their interest from this book, which makes it possible for them to accomplish further significant achievements. This book will provide a good background and extensive examples in the biological knowledge analysis and interpretation domain.

Seonho Kim
Department of Computer Science, Sogang University
1, Shinsoo-Dong, Mapo-Ku, Seoul, 121-742, Korea
E-mail: shkim@lex.yonsei.ac.kr