1 Introduction to Bioinformatics

Yi-Ping Phoebe Chen [1, 2]

[1] School of Information Technology
Faculty of Science and Technology
Deakin University, 221 Burwood Highway, VIC3125, Australia
[2] Australian Research Council (ARC) Centre in Bioinformatics

1.1 Introduction

This book is an introduction to what has come to be known as Bioinformatics and Bioinformatics Technologies. The material in this book is presented from a non-biologist's perspective, where emphasis is placed on basic concepts of Bioinformatics and technologies used to discover interesting biological data patterns unknown in large datasets. For a biologist, this book will present useful information on technologies that can be applied. Various methods that focus on the development of scalable and efficient bioinformatics technologies tools are discussed. In this chapter, you will learn how Bioinformatics is a part of the natural evolution of database technologies, why data mining, data modeling, machine learning, pattern matching, and visualization are important, and how they are defined. You will also learn about the general architecture of bioinformatics technologies and its applications. Why is it so important to understand biological problems? How can one understand a biological problem? How can one understand biological worlds from the points of view of information technology, computer science, mathematics, and commerce? These questions

are briefly answered. Furthermore, various types of biological data are discussed. This book explains the technologies that can be used for analysis, the nature of biological knowledge that can be found, and the bioinformatics tools that can be applied. Finally, challenging research issues for building bioinformatics technologies, tools, and applications of the future are also discussed.

1.2 Needs of Bioinformatics Technologies

What is bioinformatics? Why is bioinformatics important? Bioinformatics has attracted a great deal of attention from various disciplines, such as information technology, mathematics, and non-traditional biological sciences in recent years. This is due to the availability of enormous amounts of public and private biological data and the compelling need to transform biological data into useful information and knowledge. Understanding the correlations, structures, and patterns in biological data are the most important tasks in bioinformatics. The information and knowledge from these disciplines can then be wisely used for applications that cover drug discovery, genome analysis and biological control.

Bioinformatics can therefore be considered to be the combination of several scientific disciplines that include biology, biochemistry, mathematics, and computer science. It involves the use of computer technologies and statistical methods to manage and analyze a huge volume of biological data about DNA, RNA, and protein sequences, protein structures, gene expression profiles, and protein interactions.

Specifically, bioinformatics encompasses the development of databases to store and retrieve biological data, of algorithms and statistics to analyze and determine relationships in biological data, and of statistical tools to identify, interpret, and mine datasets. Figure 1.1 illustrates the underlying definition of bioinformatics (Baxevanis and Ouellette, 2001; Kuonen, 2003; Baldi and Brunak, 2001; and Westhead et al., 2002).

The field of bioinformatics plays an increasing role in the study of fundamental biological problems owing to the exponential explosion of sequence and structural information with time (Ohno-Machado et al., 2002). Figure 1.2 shows the exponential growth of GenBank.

As an example, the number of entries in a database of gene sequences in GenBank has increased from 1,765,847 to 22,318,883 in the last five years. These entries tend to double every 15 months (Benson et al., 2002).

There are two major challenging areas in bioinformatics: (1) data management and (2) knowledge discovery.

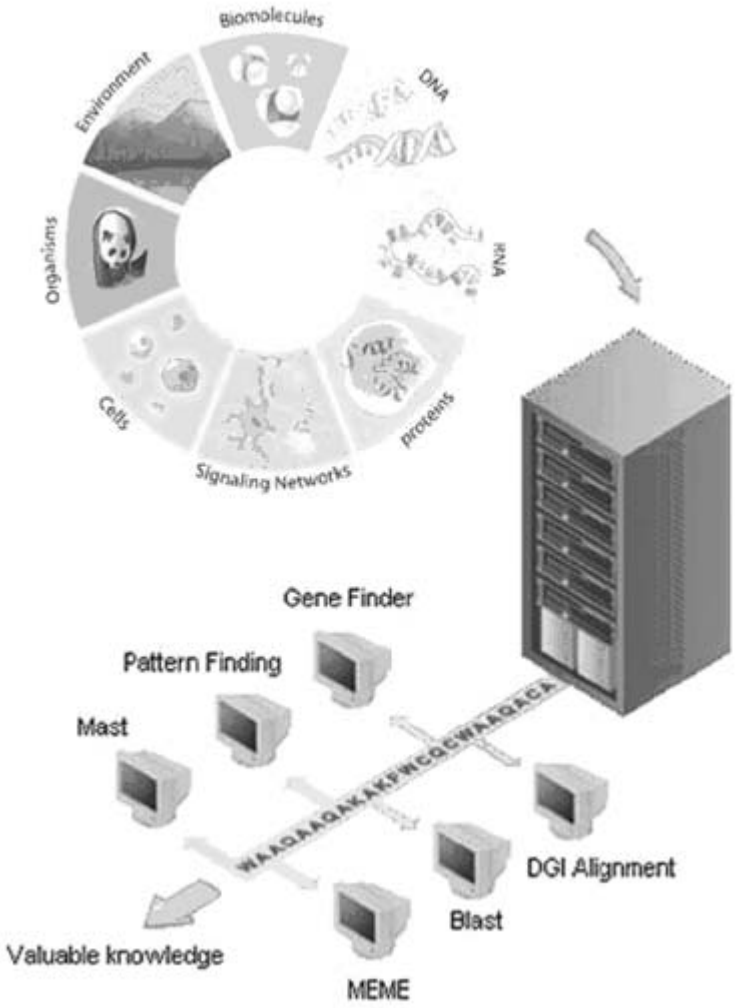

Fig. 1.1. A illustration of a bioinformatics paradigm (adapted from http://www.bioteach.ubc.ca/Bioinformatics/whatisbioinform/)

With the emergence of high-throughput technologies such as whole genome sequencing and DNA microarrays, large volumes of data are generated. The efficient management of this biological data is desirable.

A challenge to data management involves managing and integrating the existing biological databases. There are several types of databases

available to researchers in the field of biology. The most widely used among them are

- primary nucleic acid databases
  - GenBank (NCBI),
  - the Nucleotide Sequence Database (EMBL), and
  - DNA Data Bank of Japan (DDBJ)
- protein sequences databases
  - SWISS-PROT, and
  - TrEMBL
- structural databases
  - Protein Data Bank (PDB), and
  - Macromolecular Structure Database (MSD)
- literature databases
  - Medline

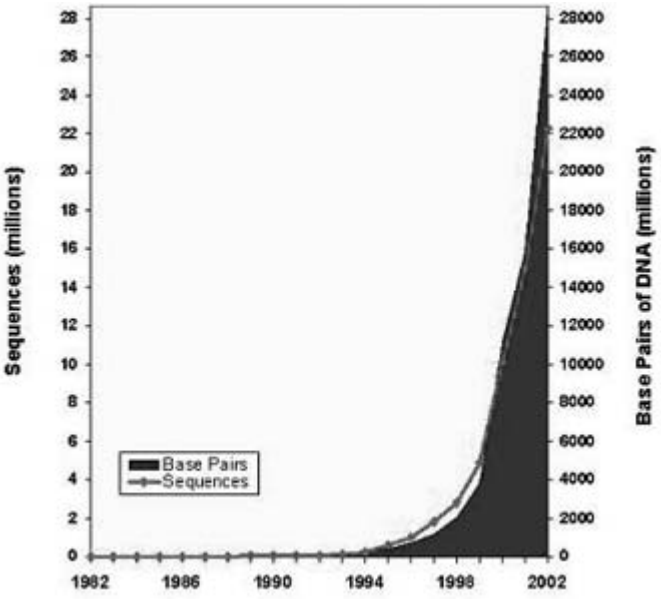

Fig. 1.2. The growth of data in GenBank (source: http://www.ncbi.nih.gov/Genbank/genbankstats.html)

However, in some situations, a single database cannot provide answers to the complex problems of biologists. Integrating or assembling information from several databases to solve problems and discover new knowledge are

other major challenges in bioinformatics (Kuonen, 2003; Ng and Wong, 2004; Wong, 2000; and Wong, 2002).

The transformation of voluminous biological data into useful information and valuable knowledge is the challenge of knowledge discovery. Identification and interpretation of interesting patterns hidden in trillions of genetic and other biological data is a critical goal of bioinformatics. This goal covers identification of useful gene structures from biological sequences, derivation of diagnostic knowledge from experimental data, and extraction of scientific information from the literature (Han and Kamber, 2001; Jagota, 2000; Narayanan et al., 2002; and Ng and Wong, 2004).

1.3 An Overview of Bioinformatics Technologies

The term "bioinformatics" has been used with different meanings by different groups of scientists and researchers (Perry, 2000). According to these researchers, it means all bioinformatics activities related to genomics that focus on chromosome mapping and sequencing, and on exploring the functions of genes, functional genomics, and structural genomics. Besides supporting genomics, information technology supports a wide range of biosciences, such as human brain science and plant architecture, and computational biology. The data are characterized by variety and heterogeneity: they are related to different organic structures, environments, and spatial scales, and derive from multiple sources. Database management, artificial intelligence, data mining, and knowledge representation can provide key solutions to the challenges posed by biological data. However, these approaches require powerful and sophisticated computational tools to provide efficient solutions to very complex problems. Exciting opportunities are emerging by integrating molecular biology components of bioinformatics with computational, physiological, morphological, taxonomic, and ecological components. Addressing these challenging issues will help the life sciences to access, retrieve, analyze, and visualize data and relationships in a collaborative work environment. Even biomedical and health informatics can benefit from bioinformatics technologies.

Bioinformatics can be viewed as naturally evolving from computer and biological sciences. This evolution has been investigated in the development of the following functionalities:

- biological data collection such as NCBI (http://www. ncbi.nih.gov/), GeneBank, DDBJ and PDB,

- biological data creation such as the human genome project, gene discovery and gene expression,
- biological databases such as EMBL, EMBI and SWISS-PROT,
- biological data management such as bioinformatics data warehousing and Sequence Retrieval Systems (SRS),
- biological data structures such as structural bioinformatics,
- biological modeling such as HMM, comparative modeling, probabilistic modeling and molecular modeling,
- biological data analysis and exploration such as bioinformatics data mining, and biological understanding such as machine learning and pattern matching and visualization of biological sequences,
- sequence analysis: sequence assembly and alignment, and
- biological processes.

Bioinformatics technology is an interdisciplinary field, a confluence of a set of technologies, as shown in Fig. 1.3. It includes database technologies, data mining, structures, process, modeling, visualization, machine learning, pattern matching, networks, and tools.

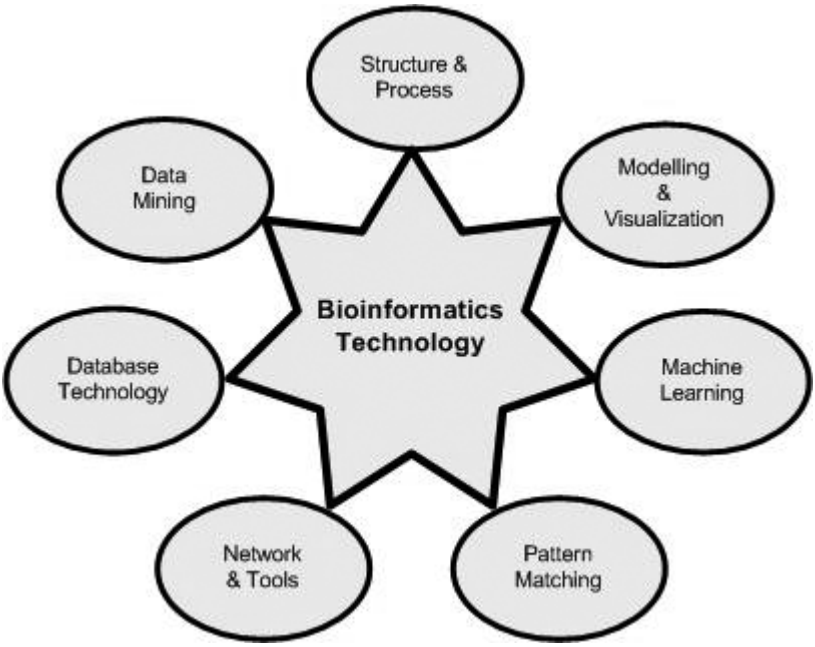

Fig. 1.3. Technologies within Bioinformatics

The existing research in bioinformatics is related to knowledge discovery, sequence analysis, structure analysis, and expression analysis. Sequence analysis is the discovery of functional and structural similarities and differences between multiple biological sequences. This can be done by comparing the new (unknown) sequence with well-studied and annotated (known) sequences. Scientists have found that two similar sequences possess the same functional role, regulatory or biochemical pathway, and protein structure. If two similar sequences are from different organisms, they are said to be homologous sequences. Finding homologous sequences is important in predicting the nature of a protein. This helps greatly in the development of new drugs, and in the performance of phylogenetic analysis. One proposed method for sequence comparison is sequence alignment. It is a procedure for base-by-base comparison of two (pairwise) or more (multiple) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. To search for an identical character or character patterns, the string matching technique is widely used. Another active research area in the field of sequence analysis is gene prediction. Gene prediction is the process of detecting meaningful signals in uncharacterized DNA sequences. Gene prediction uses homology search to acquire knowledge of the interesting information in DNA. Figure 1.4 illustrates the existing works in knowledge discovery in bioinformatics.

Structure analysis is the study of proteins and their interactions. Proteins are complex biological molecules composed of a chain of units, called amino acids, in a specific order. They are large molecules required for the structure, function, and regulation of the body's cells, tissues, and organs. Each protein has unique functions. The structures of proteins are hierarchical and consist of primary, secondary, and tertiary structures. In other words, at the molecular level, proteins can be viewed as 3D structures. The understanding of protein structures and their functions leads to new approaches for diagnosis and treatment of diseases, and the discovery of new drugs. Current research on protein structural analysis involves comparison and prediction of protein structures.

Expression analysis includes gene expression analysis and gene clustering. Basically, gene expression analysis is a study that determines the similarities or differences of genes expressed in a particular cell type or tissue. Gene expression, represented by a matrix, can be determined in two ways. First, comparing the expression profiles of genes: if the expression profiles are similar, the genes are co-regulated and functionally related. Second, by comparing the expression profiles of samples, one can consider whether genes are expressed differently. Gene clustering aims to group together genes having similar expression profiles. Genes in a specific group are co-

regulated and functionally related to each other rather than to genes in different groups. Due to the complexity and gigantic volume of biological data, the traditional computer science techniques and algorithms fail to solve the complex biological problems in the real world.

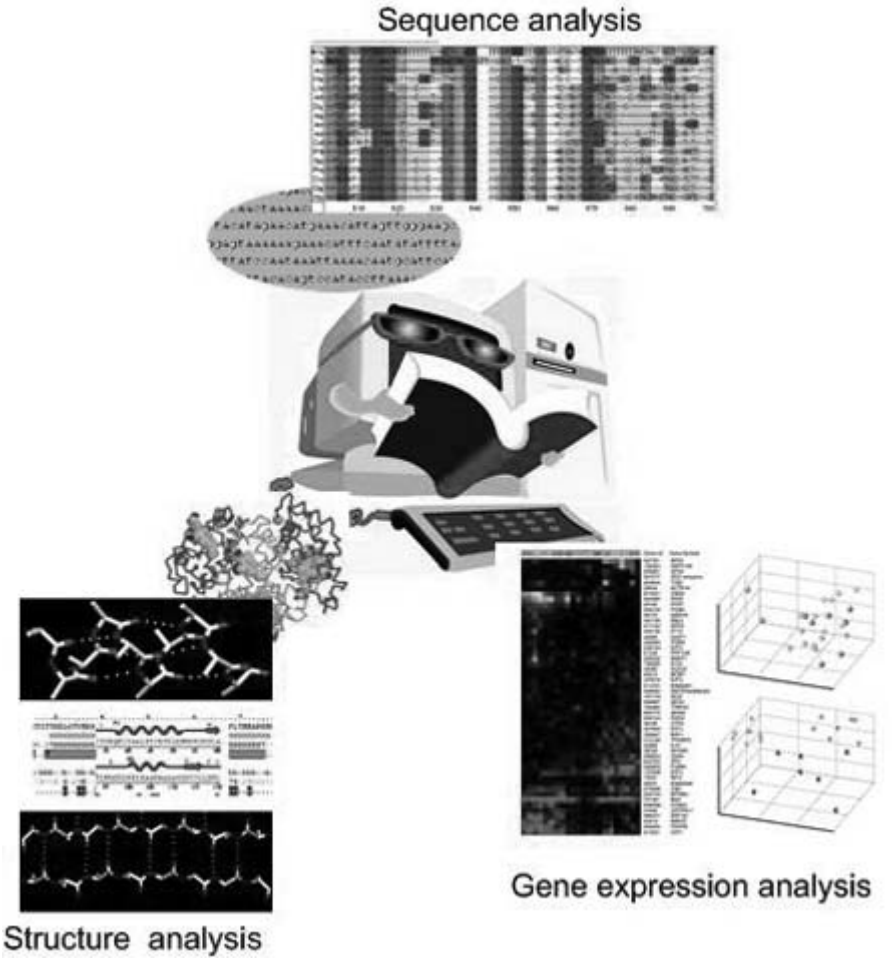

Fig. 1.4. Knowledge discovery in Bioinformatics

1.4 A Brief Discussion on the Chapters

This book covers information technology as applicable to Bioinformatics. Chapter 1 provides an overview of bioinformatics and briefly discusses the

interrelationships between the different disciplines such as biology and computer science. Furthermore, it lists in a nutshell the technologies and tools used in bioinformatics.

Chapter 2 provides an overview of structural bioinformatics. The resources of protein structures such as the Protein Data Bank (PDB), and tools and their applications are also discussed. It also covers structural classification, structure prediction, and functional assignments in structural genomics. Further, protein-protein interactions and protein-ligand interactions are clearly explained. The future of structural bioinformatics is also explored in this chapter. Chapter 2 offers a clear and concise overview that forms the foundation for Chap. 3 and part of Chap. 4.

Chapter 3 deals with the basics of database warehousing related to Bioinformatics. It dwells on the organization of bioinformatics data, and the techniques used to transform the data into meaningful information and knowledge. Data is stored in different databases located in different parts of the world, in different formats. This creates insurmountable problems for the bioinformatics community in the extraction of meaningful and reliable information. A detailed discussion is presented on data warehouse architectures and data quality to address these problems. This becomes the basis for data mining, discussed in the subsequent chapter.

Chapter 4 discusses techniques used in data mining for bioinformatics, such as biomedical data analysis, DNA data analysis, and protein data analysis. In order to discover knowledge from the vast genomic and proteomic data, we need tools to deal with the data. Pattern discovery tools and visualization tools are discussed in this chapter. A brief discussion is presented on the theory underlying DNA and protein sequences. The analytical techniques for DNA sequence comparison, gene prediction, and phylogenic analysis are subsequently explained. In the case of protein data analysis, the popular techniques, such as neural networks and HMM, and tools such as DALI and VAST, are elaborated to throw light on the secondary and tertiary protein structures. In order to mine reliable knowledge from biological data, efficient machine learning techniques are needed. These can be found in Chap. 5.

As the growth of biological data has been enormous in the last decade alone, we need to have less-time-consuming and more-reliable techniques to deal with this situation. This can be done effectively only when artificial intelligence (AI) is introduced into the processing, as exemplified in the mature engineering fields. This opens new avenues for introducing the proven AI techniques for analyzing genome and proteome. Chapter 5 deals with the major machine learning techniques, namely, artificial neural networks, genetic algorithms, and fuzzy systems. The newly evolving support vector machine is also covered in brief. Further, this chapter explains the

underlying issues of these machine learning techniques when applied to complex biological data.

Chapter 6 introduces an integrated approach, called systems biotechnology, to understand the underlying biological processes and solve the complex biological problems. The knowledge gained so far will help us to look at problems in an interrelated way. In systems biotechnology, various components, namely, experimental finding, modeling and simulation, and knowledge discovery are combined as a single system to gain insight into any biological organism. An interesting discussion can be found on the analysis of the *E. coli* genome using this approach. The chapter also discusses how a biotechnology process can be developed in a rational and systematic way. The tools necessary to implement this approach are also described in this chapter. One of the crucial components of this approach is the modeling of biological processes and data. Chapter 7 covers modeling and simulation. Chapter 7 and Chap. 9 explore all the major modeling techniques, and modeling and simulation, respectively.

Chapter 7 uses a formal language approach called Petri nets for solving the problems of biological processes. It demonstrates the effectiveness of this approach developing a software tool that can model and simulate any complex biological process using a hybrid functional Petri net with Extension (HFPNe). It is also a novel integrated approach that complements the system biotechnology approach explained in Chap. 6. A generic XML format is introduced to describe biological processes with HFPNe. The importance of visualization in the simulation of biological processes is also discussed in this chapter. This chapter covers computational modeling of biological processes with Petri net-based architectures. It also describes hybrid Petri nets and hybrid dynamic nets, hybrid functional Petri nets, implementation of HFPNe in genomic object nets, modeling of biological processes with HFPNe, and modeling from DNA to mRNA in eucaryotes and genomic object nets.

When the data to be analyzed is huge, the computational time required to analyze the data may run into days, or into months in some cases. Parallel computing alleviates this problem by making the processors efficiently use robust algorithms. Chapter 8 discusses parallel biological computing. The main components of intensive biological computing, namely, sequence assembly and sequence alignment, are discussed in this chapter. They have benefited a lot from parallel computing and will benefit more from the further research on parallel biological computing. This chapter introduces recent research on parallel sequence assembly and alignment. This chapter also provides good coverage of the main methods used in sequence assembly and sequence alignment.

Modeling plays a major role in the advancement of science and technology in any field since a good model will eventually become automated in the computational process. Bioinformatics is no exception. Chapter 9 describes modeling in bioinformatics: any representation that simulates a model of biological process. This chapter deals with important modeling approaches used in bioinformatics, namely, hidden Markov models (HMM), comparative modeling, probabilistic modeling, and molecular modeling. HMM has already taken a strong root in bioinformatics after its widespread use in speech recognition. Comparative modeling does great service to drug discovery as it relates the structure and functions of the protein as well as the gene. This chapter provides a detailed discussion on comparative modeling of proteins and genes. The theory of probability has made inroads into bioinformatics as well. The main probabilistic modeling techniques, namely, (1) Bayesian networks, (2) stochastic context-free grammars, and (3) probabilistic Boolean networks, are discussed. In order to understand the biological processes properly, knowledge of the molecule and molecular interactions are very much needed so that the projected functionality of any new drug under development can be verified. Molecular modeling provides such knowledge in the form of the molecular structure in terms of structural attributes such as bond angle, bond length, torsion angle, and potential energy. Further, the simulation that comes out of this modeling paves a way for further research into the intricacies of the molecular dynamics.

In all the chapters we have so far discussed the techniques or modeling to retrieve information from the molecular sequences. Chapter 10 discusses how we can locate a particular segment of the sequence, known as the motif or pattern, which is responsible for a particular manifestation such as a disease. This chapter deals with pattern matching and motif discovery. The major computational approaches used to find motifs are clearly described.

Knowledge of the spatial geometry sheds light on molecular structure. Fractal theory deals with such spatial geometry. It provides a mathematical formalism to describe any complex spatial and dynamic structure. It has been successfully applied to the study of many problems in science and engineering. Application of fractal theory on the structure of DNA and proteins is expected to solve complex problems that seem incomprehensible at the moment. Chapter 11 presents some tools built on the theory of fractal geometry that may play a useful role in solving biological problems. This chapter discusses the popular multifractal analysis used to characterize the spatial heterogeneity of both theoretical and experimental fractal patterns in DNA and protein sequences.

As we are aware, DNA contains numerous genes. The functions of each gene are not fully explored. Using traditional methods, several experiments

need to be conducted to find out the functions of a single gene alone. However, modern technologies create opportunities to conduct experiments to find out the locations and the functions of genes simultaneously, using a technique called a microarray. Thousands of DNA samples are coated with glass or nylon in the form of a two-dimensional array, and encapsulated in a microchip for spectroscopic analysis. Chapter 12 deals with the microarray technique and microarray data analysis. It also explains knowledge discovery, data mining, clustering, and classification. Techniques on protein information resources and DNA sequence analysis are also covered.

References

Attwood, T.K. and Parry-Smith, D.J. (1999) Introduction to Bioinformatics, Prentice Hall.

Baldi, P. and Brunak, S. (2001) "Bioinformatics The Machine Learning Approach", The MIT Press.

Baxevanis, A.D. and Ouellette, B.F.F. (2001) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2nd ed. New York: Wiley-Interscience, 2001.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2002) "GenBank," Nucleic Acids Res., vol. 28, pp. 15-18, 2002.

Han, J. and Kamber, M. (2001) "Data Mining: Concepts and Techniques". San Francisco, Calif.: Morgan Kaufmann, 2001.

Jagota, A. (2000) "Data Analysis and Classification for Bioinformatics". California: Bay Press, 2000.

Kuonen, D. (2003) "Challenges in bioinformatics for statistical data miners", Bulletin of the Swiss Statistical Society, vol. 46, pp. 10-17.

Narayanan, A., Keedwell, E.C. and Olsson, B. (2002) "Artificial Intelligence Techniques for Bioinformatics," Applied Bioinformatics, vol. 1, pp. 191-222.

Ng, S.K. and Wong, L. (2004) "Accomplishments and challenges in bioinformatics", IT Professional, 2004, vol 6, issue: 1, pp. 44- 50.

Ohno-Machado, L., Vinterbo, S., Weber, G. (2002) "Classification of gene expression data using fuzzy logic", Journal of Intelligent and Fuzzy Systems, vol. 12, pp. 19-24.

Perry, L.M. (2000) "Focus on interactions with bioinformatics", Journal of the American Medical Informatics Association, vol. 7, no. 5, pp. 431-438.

Westhead, D.R., Parish, J.H. and Twyman, R.M. (2002) "Bioinformatics", Instant Notes in Bioinformatics, BIOS Scientific Publishing.

Wong, L. (2002) "Technologies for integrating biological data", Briefings in Bioinformatics, vol. 3, pp. 389-404.

Wong, L. (2000) “Kleisli, a functional query system”, Journal of Functional Programming, vol. 10, pp. 19-56.