

Chapter 1

Introduction to Bioinformatics

1.1 What Is Bioinformatics

As we enter the information age, we witness the impact of computers and computation in almost every corner of our lives. Many people in the world retrieve and broadcast information through the Internet. The weather forecast is made through extensive computation on supercomputers. Stocks are traded electronically. Airplanes are designed completely on computers before the first component is ever manufactured. We also witness substantial impact of computers and computation on biological and medical research, and this impact led to the birth of bioinformatics.

Although bioinformatics is a popular term in science and technology, there is no consensus for its definition. As a new field, its precise definition will take many years to finalize. A current semi-official definition for bioinformatics by the US National Institutes of Health (NIH) is “Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, represent, describe, store, analyze, or visualize such data” (<http://www.bisti.nih.gov/>). A related field, computational biology, is defined by NIH as “the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems”. From these definitions, bioinformatics is focused on technology (engineering) for developing

tools and infrastructure, while computational biology is more about science (biology) to generate hypotheses in understanding nature.

Although the distinction between bioinformatics and computational biology is made by NIH and others, there is no doubt that the two fields are tightly coupled. Hence, the terms bioinformatics and computational biology are sometimes used interchangeably. For example, the definition of bioinformatics by Luscombe *et al.* [2001] includes some scope of computational biology specified by NIH, but restricts itself to the biomolecular aspect: “bioinformatics is conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying “informatics” techniques (derived from disciplines such as applied math, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale.”

Bioinformatics is deeply rooted in three traditional disciplines, i.e., biology, computer science, and statistics. Both biology and computer science often claim bioinformatics as a sub-discipline. Furthermore, bioinformatics has strong ties to physics, biophysics, mathematics, chemistry, and engineering. On the other hand, bioinformatics is becoming an independent discipline by itself, with its own theoretical foundations, analytical approaches, and computational techniques. This emergence is similar to biophysics, which evolved from an interdisciplinary field between biology and physics to an integral science.

1.2 A Brief History of Bioinformatics

Although bioinformatics is a new term developed in the early 1990s, bioinformatics research started before 1970. Over the past four decades, bioinformatics emerged gradually from a hardly noticeable area to a mainstream discipline in science. You can find a comprehensive historical perspective of bioinformatics in the review by Ouzounis and Valencia [2003]. Here, we highlight some major milestones that define today’s bioinformatics. If you are unfamiliar with some of the biological terms, you can find related materials in Appendices I and II.

In the 1960s, a number of key contributions in investigating biomolecular evolution paved the way for applying computers in studying biological sequences. Zuckerkandl and Pauling [1965] pioneered the use of biological sequences in evolutionary studies, which laid the theoretical foundation for computational studies of evolutionary patterns in genes and proteins. Fitch and Margoliash [1967] developed computational methods to build a tree structure (called “phylogenetic tree”) from gene sequences for understanding gene evolution. Margaret Dayhoff and her coworkers developed a scoring method (called a “mutation matrix”) for comparing protein sequences, and created computerized protein sequence databases for biomolecular evolution [Dayhoff *et al.*, 1965]. Because of her contribution, Dayhoff is regarded as a founder of the field of bioinformatics.

In the 1970s, a series of theoretical and computational studies opened new doors for bioinformatics research in diverse biological problems. Needleman and Wunsch [1970] published the first efficient algorithm for comparing two biological sequences based on dynamic programming. Lee and Richards [1971] provided a method for computing the geometry of protein three-dimensional structure. Chou and Fasman [1974] proposed a method for predicting protein secondary structures from a protein sequence. A few laboratories started simulation of protein dynamics and protein folding processes [Levitt and Warshel, 1975; Tanaka and Scheraga, 1975; Karplus and Weaver, 1976; Hagler and Honig, 1978]. Furthermore, RNA structure predictions emerged [Tinoco *et al.*, 1971; Waterman and Smith, 1978].

In the 1980s, various bioinformatics algorithms were significantly improved and bioinformatics tools became more sophisticated. In 1981, the Smith-Waterman algorithm for aligning two biological sequences was published [Smith and Waterman, 1981]. Although this algorithm is based on the one by Needleman and Wunsch [1970], the improvement allowed a comparison between parts of one sequence and parts of another sequence (which is called “local alignment”). This paved the way for large-scale sequence comparison and search. Because of this development and other contributions, Michael Waterman is regarded as another founder of bioinformatics. FASTA [Lipman and Pearson, 1985] was an early program for fast sequence similarity search in a database.

Feng and Doolittle [1987] developed a successful method to compare a group of sequences simultaneously (which is termed as “multiple-sequence alignment”). A number of systematic approaches for building phylogenetic trees were published, among which PHYLIP [Felsenstein, 1989] became a popular package. Kuntz *et al.* [1982] pioneered a method for predicting protein-ligand docking conformation. Computational methods for predicting genes from a DNA sequence were proposed [Shepherd, 1981; Fickett, 1982; Staden and McLachlan, 1982]. With these developments, the importance of bioinformatics research was recognized. Particularly, in 1988, the National Center for Biotechnology Information (NCBI) in the US was created to handle various bioinformatics issues from data distribution to data analysis.

The golden age of bioinformatics started in 1990s. This boom was mainly due to the Human Genome Project, which officially started in 1990. The goal of this project was to determine the sequence of the entire human genome. Genomic sequencing has opened a new avenue to study biological systems on large scales, setting the stage for generating many other high-throughput data. The new techniques for studying biology in large scale raised various new challenges for bioinformatics. Phil Green and his colleagues addressed the computational problem of identifying nucleotides from image data of a sequencer, a process referred to as “base calling” [Ewing *et al.*, 1998]. A widely used method for genome sequencing is the “shotgun” approach, where bioinformatics is required to assemble short, overlapping pieces of DNA sequences into a long, coherent sequence. Green [2002] and Myers [1995] developed methods for solving this problem, which was a major contribution to the Human Genome Project.

In 1990, the exponential growth of biomolecular data clearly showed the need for interpreting, managing and mining these data. Various bioinformatics databases, such as GenBank (<http://www.ncbi.nlm.nih.gov/Genbank>), a database for biological sequences, became essential to biomedical research. Many bioinformatics algorithms led to sophisticated computer packages, with user-friendly interfaces. Meanwhile, computers became faster and cheaper and the Internet provided a major platform for accessing bioinformatics tools and databases. Many experimental biologists started

to use various bioinformatics packages, especially through Web interfaces [Xu *et al.*, 2000; Rhee *et al.*, 2006]. A number of popular software packages and servers developed in the 1990s are widely used, as indicated by their large numbers of citations (see Table 1.1). The sequence comparison tool BLAST [Altschul *et al.*, 1990] became a household name to biologists. It is also the most popular tool among all the computational tools that have ever been developed since the birth of the computer, with its defining paper as the most cited reference in the scientific history. The widespread use of bioinformatics applications has had an enormous impact on research in biology and medicine.

Table 1.1. Popular Bioinformatics Packages

Name	Functionality	URL	Reference	Citations
BLAST	Pairwise sequence alignment	http://www.ncbi.nlm.nih.gov/BLAST/	Altschul <i>et al.</i> , 1990	20,495
CLUSTAL-W	Multiple sequence alignment	http://www.ebi.ac.uk/clustalw/	Thompson <i>et al.</i> , 1994	18,837
SignalP	Signal peptide prediction	http://www.cbs.dtu.dk/services/SignalP/	Nielsen <i>et al.</i> , 1997	3002
DALI	Protein structure comparison	http://www.ebi.ac.uk/dali/	Holm and Sander, 1993	g
MODELLER	Protein tertiary structure prediction	http://www.salilab.org/modeller/	Sali and Blundell, 1993	1817
PHD	Protein secondary structure prediction	http://www.predictprote.in.org/	Rost and Sander, 1993	1795
SEQUEST	Protein identification using mass-spec data	http://fields.scripps.edu/sequist	Eng <i>et al.</i> , 1994	1324
MFOLD	RNA secondary structure prediction	http://www.bioinfo.rpi.edu/applications/mfold/	Mathews <i>et al.</i> , 1999	1228
PHRED	DNA sequencing	http://www.phrap.org/	Ewing <i>et al.</i> , 1998	1162
GENESCAN	Gene identification in DNA	http://genes.mit.edu/GENSCAN.html	Burge and Karlin, 1997	1139

Note: The number of journal citations was based on the “ISI Web of Knowledge” (<http://nadc.isiknowledge.com>) on August 4, 2006.

Coming into the new millennium, bioinformatics became a very active research field as modern biology quickly evolves. The availability of genomic sequences enabled a number of new high-throughput measurement technologies, which expanded genome-scale studies from

sequence-level information to higher-level functions. For example, microarrays are powerful tools for systematic measurement of large-scale gene-expression data under varying experimental conditions or over a time course. Various experimental methods can generate different types of protein-protein interaction information [Chen and Xu, 2003]. Each new experimental technique for large-scale biomolecular measurement often requires new development in data interpretation and analysis. In microarrays, extensive studies have been conducted in image processing, statistical analysis, and clustering [Speed, 2003]. In recent years, new sequencing techniques [Service, 2006], such as the 454 sequencer [Margulies *et al.*, 2005], were developed to reduce the cost of sequencing. The bioinformatics challenges in these sequencing technologies are numerous in terms of experimental design, data interpretation, and data integration.

In recent years, systems biology emerged as a field which integrates experimental, theoretical, and computational techniques to study biological organisms at multiple levels as a system instead of individual components [Alon, 2006]. A systems approach brings renewed hope for solving some long-standing biomedical problems, especially various complicated diseases such as cancer and diabetes. Bioinformatics is a key component in systems biology, bringing heterogeneous data together for analysis, modeling and design [Kriete and Eils, 2005]. For example, bioinformatics can be used to predict a biomolecular network as a large-scale system [Palsson, 2006]. In addition, it also helps to fuse and integrate a wide spectrum of high-throughput data, including biological sequences, gene expression levels, protein interactions, small RNA regulation [Washietl *et al.*, 2005], epigenomics data [Model *et al.*, 2001], and metabolomic data [Steuer *et al.*, 2003].

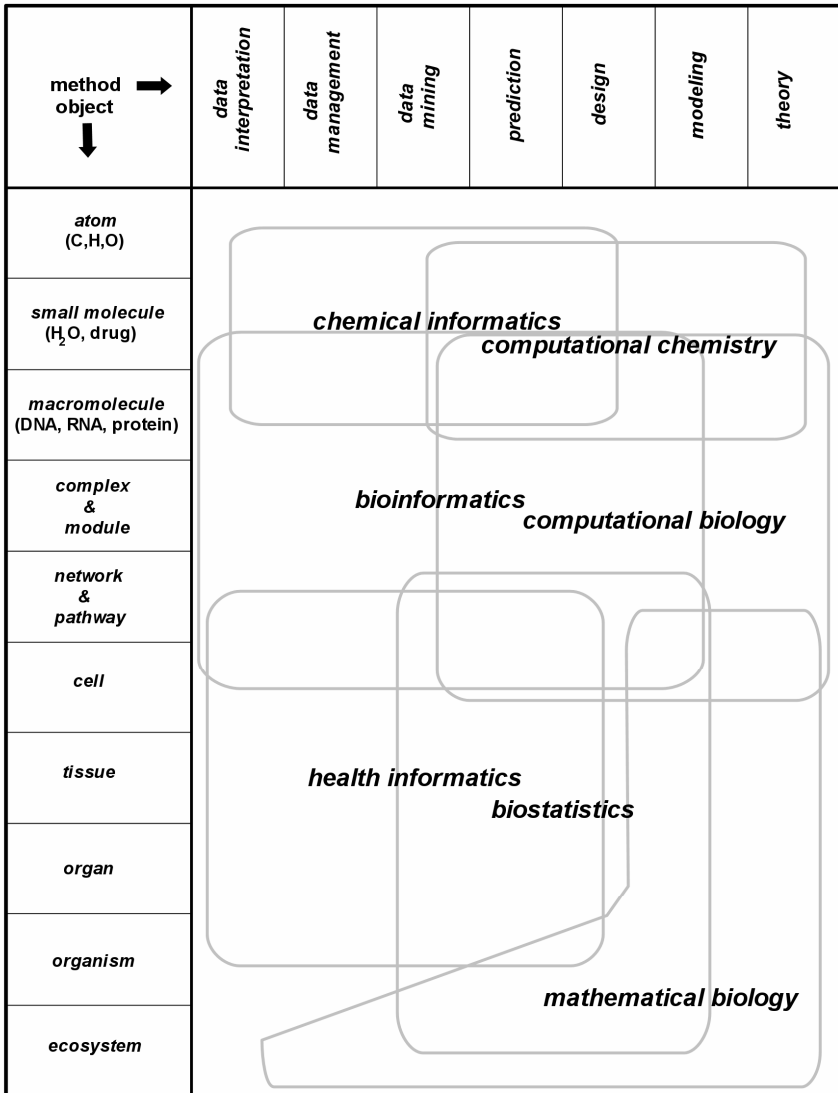


Figure 1.1 Our view of the scope of bioinformatics and related areas in a matrix of biological objects and computational approaches.

1.3 Scope of Bioinformatics

Regardless of its definition, the scope of bioinformatics is extremely broad and is rapidly changing, particularly in recent years. Although bioinformatics theoretically could address all bio-related issues, the current scope of bioinformatics is mainly at the biomolecular level (see Figure 1.1), particularly on macromolecules (DNA, RNA, and proteins), biological complexes/modules involving a group of genes/proteins, and biomolecular networks/pathways that control various interactions among genes/proteins. The roles of bioinformatics in modern biology are summarized in Figure 1.2. More specifically, bioinformatics targets the following major computational issues and methods:

1) Data interpretation in high-throughput technologies

Various high-throughput technologies became the driving force of modern biology. These technologies include sequencers for DNA sequencing, mass spectrometers for protein identification, microarrays for gene expression profiling, etc. Typically, the initial outputs of these technologies are images and spectra, which are often huge in size and noisy in data quality. Computational methods are required to process these data; thus, bioinformatics plays an important role to automate the interpretation of the images and spectra and to convert them into numerical values.

2) Data management and computational infrastructure

Given the size and complexity of biological data, creation and maintenance of databases of biological information are essential to modern biology. Biological sequences and their annotations comprise the majority of such databases, while many other types of databases for microarray gene expression, protein structures, etc. are expanding quickly. Bioinformatics handles the design of these databases for data storage, update, and retrieval. In many cases, a Web interface is provided for data access, together with some back-end engines for data analyses (see Appendix II for examples). Sometimes graphical tools or plug-ins are provided for data visualization. Furthermore, some biological databases may connect to experimental instruments for real-time data collection using a tracking system, such as a Laboratory Information Management Systems (LIMS) [Paszko and Turner, 2001].

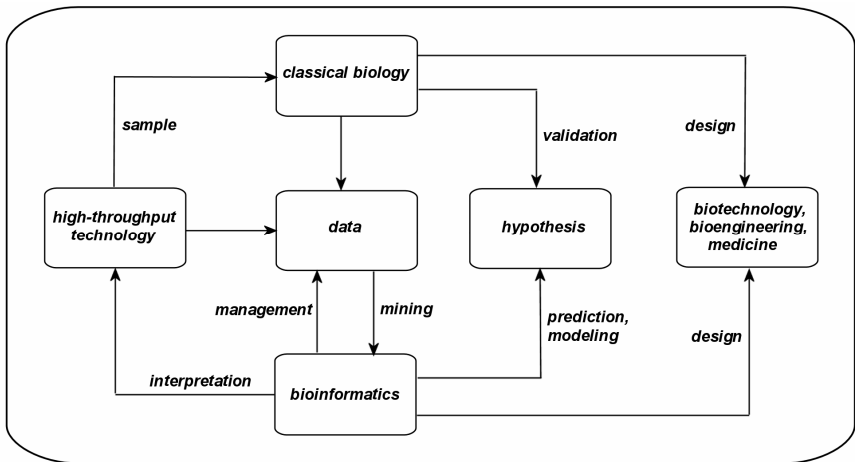


Figure 1.2 Roles of bioinformatics in modern biology.

3) Discovery from data mining

A demanding task for bioinformatics is to extract useful biological information and patterns from noisy data produced by high-throughput technologies. For example, one can compare sequences of multiple genomes to identify interesting evolutionary patterns. Analyzing microarray data can lead to the discovery of the genes that are associated with a particular disease. Mining biomedical literature can lead to automated identification of possible gene-gene associations. We will address microarray data mining extensively in Chapter 5 of this book.

4) Prediction

Bioinformatics is often used to predict biological information. In particular, from a protein sequence alone, one can predict protein secondary structure, protein localization (in a compartment of a cell), protein function, etc. We will discuss protein secondary structure prediction in detail in Chapter 4 of this book. Data mining techniques can be applied in bioinformatics predictions. For example, in protein structure prediction, one can mine known protein structures with similar protein sequences to a query protein and then construct a structural model based on the known structures, in a process called “homology modeling” [Sali and Blundell, 1993]. Bioinformatics prediction is becoming an integral part of modern biology through an iterative process of hypothesis generation and experimental validation.

5) Computational design

Bioinformatics is widely used as a design tool in medicine and bioengineering. One can use bioinformatics in protein structure-based or gene-based drug discovery and development. It can help design a delivery using a certain combination of drugs and at a certain schedule to achieve maximum performance (e.g., for AIDS treatment). It can also suggest mutations of a gene for achieving certain biological properties in a genetically modified species. For example, it is possible to suggest a mutation of a gene to achieve drought resistance in soybeans.

6) Modeling

Modeling of biological systems and processes often adds value to the available biological data. A well established area in bioinformatics is protein structure modeling [Xu *et al.*, 2006]. One can model various aspects of a protein structure, including geometry, energetics, and dynamics. For example, a useful modeling technique [Nicholls *et al.*, 1991] calculates and visualizes the electrostatic field of a protein structure. One can also model a neural system using differential equations, with parameters fitting some experimental data. In many cases, these parameters cannot be measured directly. Modeling can be used to interpret experimental results and to generate new hypotheses. In recent years, an entire cell has been modeled. For example, E-Cell (<http://www.e-cell.org/>) attempts to model and reconstruct biological phenomena computationally and perform whole cell simulations.

As bioinformatics expands its scope, a number of areas emerge as sub-disciplines. Each of the sub-disciplines has its own special methods and techniques. Structural bioinformatics focuses on the computational analysis and prediction of macromolecular structure (especially protein structure). Computational proteomics handles management and analysis of proteomics data for protein identification and protein interaction determination. Computational systems biology addresses algorithm and application development for systems biology. On the application side, sub-disciplines focus on the application of bioinformatics in different biological subjects. For example, immunoinformatics models immunological components for better understanding immune functions. Pharmacoinformatics deals with drug discovery using bioinformatics approaches. Agroinformatics (agricultural informatics) specializes in the bioinformatics that deals with plants and domestic animals.

In addition to expansion in various sub-disciplines, there are overarching issues for bioinformatics. One of them is bioinformatics standards. As almost all the analyses in bioinformatics are large scale, automated processing without extensive manual interruption is essential. For different tools and databases to communicate with each other, some standards need to be established. One of the efforts is ontology, which is a set of controlled vocabularies. We will have a thorough discussion of

ontologies in Chapter 3. An infrastructure to facilitate interactions between databases and servers is the semantic web, which creates a universal mechanism for information exchange and reuse in a machine-interpretable way across application, organization, and community boundaries [Neumann, 2005].

Bioinformatics has a number of related fields (in addition to computational biology), as illustrated in Figure 1.1. At the small molecule level, cheminformatics applies information technology in identification and optimization of drug leads, while computational chemistry employs quantitative methods for calculating molecular properties or simulating molecular behavior. At the macroscopic scale, health informatics (or medical informatics) addresses computational development for improving communication, understanding and management of medical information and practice; biostatistics applies statistical techniques in health-related fields, such as medicine, biology, and public health; and mathematical biology (or biomathematics) uses theoretical and numerical methods and tools to model biological systems and processes. The areas shown in Figure 1.1 overlap with each other and their scopes have been historically defined, although different researchers have different views. Among all the areas, a unique hallmark of bioinformatics is its emphasis on the development of computational tools and infrastructures driven by the need of users instead of the developers.

1.4 Major Challenges in Bioinformatics

As massive biological data have become a fundamentally important resource during discovery of new biological knowledge, a key task for bioinformatics is to identify meaningful information (or statistically significant patterns) from data and correlate such information with biological knowledge. However, such a task is highly challenging in many cases: (1) the data size is large with high dimensionality, with a complexity much higher than those typically handled by traditional computational sciences; (2) the information-rich data are heterogeneous in nature, noisy, and incomplete, as well as containing misleading outliers; and (3) biological systems, due to adaptability, evolution, redundancy, robustness, and emergence, are extremely complex. Many biological data are generated by biological processes which are not well understood. Interpretation of such data requires discovery of convoluted relationships hidden in the data. Due to these challenges, the accuracy of prediction or the information mined from a database is often not satisfactory. It is clear that there is much room for further improvement and development, which require novel theoretical frameworks and computational techniques.

Other than the technical challenges, human factors are also important. Given the scope of bioinformatics, it is unlikely for a single person to have deep understanding in relevant fields of computer science, biology, and statistics. Inevitably a researcher may not have a complete view or knowledge to solve a particular problem. In most cases collaborations are needed. However, overcoming “language” barriers among researchers from different backgrounds is often demanding. Currently, a majority of experimental biologists are not familiar with concepts, methods and tools available or emerging in bioinformatics. Computational researchers often do not understand biology in depth. More communication among different disciplines is essential for bioinformatics research.

1.5 Bioinformatics and Computer Science

The challenges in bioinformatics have resulted in a wide range of studies from computer sciences. Almost all available computer science techniques have been applied in bioinformatics. The following are some most notable applications of computational and statistical methods in bioinformatics:

- 1) Dynamic programming
- 2) Neural networks
- 3) Hidden Markov Models
- 4) Hypothesis test
- 5) Bayesian statistics
- 6) Clustering
- 7) Sampling search (Gibbs, Monte Carlo, etc)
- 8) Maximum likelihood methods
- 9) Information theory
- 10) Support Vector Machines

Fuzzy set theory has been used in bioinformatics, but to a much less extent than any of the methods above. We believe there is much higher potential for fuzzy set theory in bioinformatics, and hence, the focus of this book.

Not only does computer science provide techniques for bioinformatics, bioinformatics is also a new driver of computer science. Better hardware (supercomputers) is often demanded by bioinformatics applications. New data representation and new algorithm development fuel active research in computer science. Bioinformatics may also inspire new theoretical frameworks for computer science. Traditionally, a number of computational techniques came from biological concepts, such as neural networks, genetic algorithms, automata, and fuzzy set theory. In recent years, DNA computing is being developed to use DNA and biochemical reactions, instead of the traditional silicon-based computer chips to solve computational problems [Adleman, 2004]. Ant colony optimization mimics the behavior of ants in finding paths from the colony to food and uses a probabilistic technique for solving

computational problems [Dorigo and Stützle, 2004]. Meanwhile particle swarm optimization [Eberhart and Kennedy, 1995] imitates social communication (say, among insects) to produce cooperation in groups of potential solutions in hunting for very good answers to highly complex problems. More computer science techniques will be developed with the research of bioinformatics.

If the reader wishes to know more about bioinformatics, we suggest some related books for reference [Jiang *et al.*, 2002; Claverie, 2003; Jones and Pevzner, 2004; Lesk, 2005] and review articles [Luscombe *et al.*, 2001; Ouzounis and Valencia, 2003; Kanehisa and Bork, 2003; Bonetta, 2004; Rhee *et al.*, 2006], as well as the Internet resources listed in Appendix II.