SECTION ONE

Introduction and Biological Databases

CHAPTER ONE

Introduction

Quantitation and quantitative tools are indispensable in modern biology. Most biological research involves application of some type of mathematical, statistical, or computational tools to help synthesize recorded data and integrate various types of information in the process of answering a particular biological question. For example, enumeration and statistics are required for assessing everyday laboratory experiments, such as making serial dilutions of a solution or counting bacterial colonies, phage plaques, or trees and animals in the natural environment. A classic example in the history of genetics is by Gregor Mendel and Thomas Morgan, who, by simply counting genetic variations of plants and fruit flies, were able to discover the principles of genetic inheritance. More dedicated use of quantitative tools may involve using calculus to predict the growth rate of a human population or to establish a kinetic model for enzyme catalysis. For very sophisticated uses of quantitative tools, one may find application of the "game theory" to model animal behavior and evolution, or the use of millions of nonlinear partial differential equations to model cardiac blood flow. Whether the application is simple or complex, subtle or explicit, it is clear that mathematical and computational tools have become an integral part of modern-day biological research. However, none of these examples of quantitative tool use in biology could be considered to be part of bioinformatics, which is also quantitative in nature. To help the reader understand the difference between bioinformatics and other elements of quantitative biology, we provide a detailed explanation of what is bioinformatics in the following sections.

Bioinformatics, which will be more clearly defined below, is the discipline of quantitative analysis of information relating to biological macromolecules with the aid of computers. The development of bioinformatics as a field is the result of advances in both molecular biology and computer science over the past 30–40 years. Although these developments are not described in detail here, understanding the history of this discipline is helpful in obtaining a broader insight into current bioinformatics research. A succinct chronological summary of the landmark events that have had major impacts on the development of bioinformatics is presented here to provide context.

The earliest bioinformatics efforts can be traced back to the 1960s, although the word *bioinformatics* did not exist then. Probably, the first major bioinformatics project was undertaken by Margaret Dayhoff in 1965, who developed a first protein sequence database called *Atlas of Protein Sequence and Structure*. Subsequently, in the early 1970s, the Brookhaven National Laboratory established the Protein Data Bank for archiving three-dimensional protein structures. At its onset, the database stored less

than a dozen protein structures, compared to more than 30,000 structures today. The first sequence alignment algorithm was developed by Needleman and Wunsch in 1970. This was a fundamental step in the development of the field of bioinformatics, which paved the way for the routine sequence comparisons and database searching practiced by modern biologists. The first protein structure prediction algorithm was developed by Chou and Fasman in 1974. Though it is rather rudimentary by today's standard, it pioneered a series of developments in protein structure prediction. The 1980s saw the establishment of GenBank and the development of fast database searching algorithms such as FASTA by William Pearson and BLAST by Stephen Altschul and coworkers. The start of the human genome project in the late 1980s provided a major boost for the development of bioinformatics. The development and the increasingly widespread use of the Internet in the 1990s made instant access to, and exchange and dissemination of, biological data possible.

These are only the major milestones in the establishment of this new field. The fundamental reason that bioinformatics gained prominence as a discipline was the advancement of genome studies that produced unprecedented amounts of biological data. The explosion of genomic sequence information generated a sudden demand for efficient computational tools to manage and analyze the data. The development of these computational tools depended on knowledge generated from a wide range of disciplines including mathematics, statistics, computer science, information technology, and molecular biology. The merger of these disciplines created an information-oriented field in biology, which is now known as *bioinformatics*.

WHAT IS BIOINFORMATICS?

Bioinformatics is an interdisciplinary research area at the interface between computer science and biological science. A variety of definitions exist in the literature and on the world wide web; some are more inclusive than others. Here, we adopt the definition proposed by Luscombe et al. in defining bioinformatics as a union of biology and informatics: *bioinformatics* involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins. The emphasis here is on the use of computers because most of the tasks in genomic data analysis are highly repetitive or mathematically complex. The use of computers is absolutely indispensable in mining genomes for information gathering and knowledge building.

Bioinformatics differs from a related field known as *computational biology*. Bioinformatics is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products and is often considered *computational molecular biology*. However, computational biology encompasses all biological areas that involve computation. For example, mathematical modeling of ecosystems, population dynamics, application of the game theory in behavioral studies, and phylogenetic construction using fossil records all employ computational tools, but do not necessarily involve biological macromolecules.

Beside this distinction, it is worth noting that there are other views of how the two terms relate. For example, one version defines *bioinformatics* as the development and application of computational tools in managing *all kinds* of biological data, whereas *computational biology* is more confined to the theoretical development of algorithms used for bioinformatics. The confusion at present over definition may partly reflect the nature of this vibrant and quickly evolving new field.

GOALS

The ultimate goal of bioinformatics is to better understand a living cell and how it functions at the molecular level. By analyzing raw molecular sequence and structural data, bioinformatics research can generate new insights and provide a "global" perspective of the cell. The reason that the functions of a cell can be better understood by analyzing sequence data is ultimately because the flow of genetic information is dictated by the "central dogma" of biology in which DNA is transcribed to RNA, which is translated to proteins. Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and sometimes structural approaches has proved to be a fruitful endeavor.

SCOPE

Bioinformatics consists of two subfields: the development of computational tools and databases and the application of these tools and databases in generating biological knowledge to better understand living systems. These two subfields are complementary to each other. The tool development includes writing software for sequence, structural, and functional analysis, as well as the construction and curating of biological databases. These tools are used in three areas of genomic and molecular biological research: molecular sequence analysis, molecular structural analysis, and molecular functional analysis. The analyses of biological data often generate new problems and challenges that in turn spur the development of new and better computational tools.

The areas of sequence analysis include sequence alignment, sequence database searching, motif and pattern discovery, gene and promoter finding, reconstruction of evolutionary relationships, and genome assembly and comparison. Structural analyses include protein and nucleic acid structure analysis, comparison, classification, and prediction. The functional analyses include gene expression profiling, protein–protein interaction prediction, protein subcellular localization prediction, metabolic pathway reconstruction, and simulation (Fig. 1.1).

The three aspects of bioinformatics analysis are not isolated but often interact to produce integrated results (see Fig. 1.1). For example, protein structure prediction depends on sequence alignment data; clustering of gene expression profiles requires the use of phylogenetic tree construction methods derived in sequence analysis. Sequence-based promoter prediction is related to functional analysis of

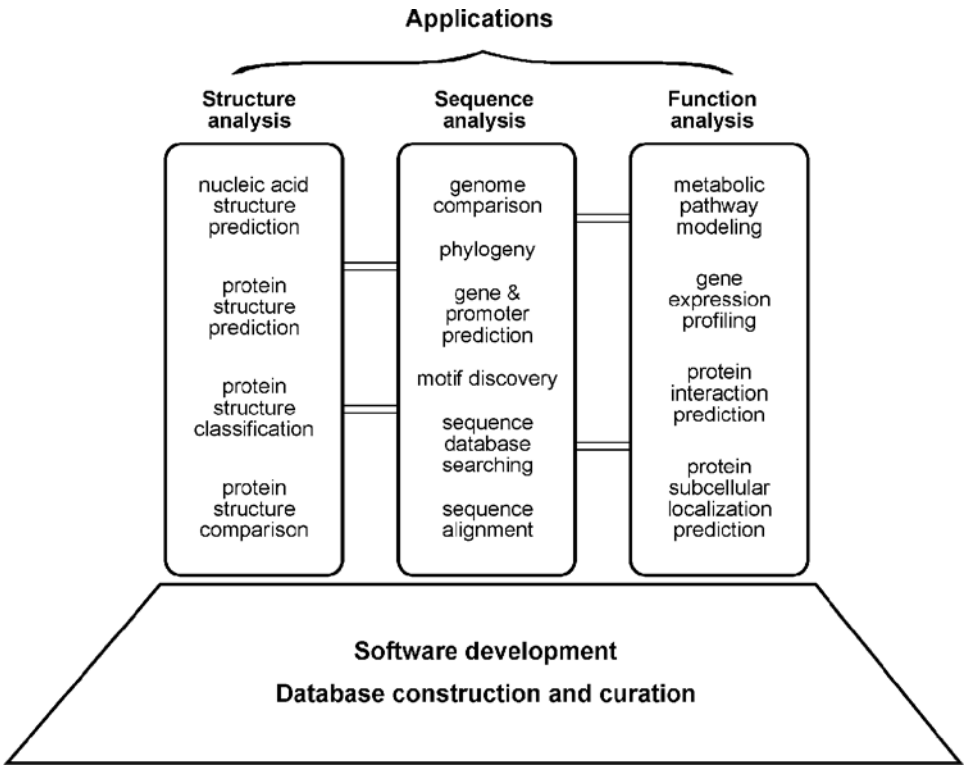

Figure 1.1: Overview of various subfields of bioinformatics. Biocomputing tool development is at the foundation of all bioinformatics analysis. The applications of the tools fall into three areas: sequence analysis, structure analysis, and function analysis. There are intrinsic connections between different areas of analyses represented by bars between the boxes.

coexpressed genes. Gene annotation involves a number of activities, which include distinction between coding and noncoding sequences, identification of translated protein sequences, and determination of the gene's evolutionary relationship with other known genes; prediction of its cellular functions employs tools from all three groups of the analyses.

APPLICATIONS

Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences. It has applications, for example, in knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology. Computational studies of protein–ligand interactions provide a rational basis for the rapid identification of novel leads for synthetic drugs. Knowledge of the three-dimensional structures of proteins allows molecules to be designed that are capable of binding to the receptor site of a target protein with great affinity and specificity. This informatics-based approach

significantly reduces the time and cost necessary to develop drugs with higher potency, fewer side effects, and less toxicity than using the traditional trial-and-error approach. In forensics, results from molecular phylogenetic analysis have been accepted as evidence in criminal courts. Some sophisticated Bayesian statistics and likelihood-based methods for analysis of DNA have been applied in the analysis of forensic identity. It is worth mentioning that genomics and bioinformtics are now poised to revolutionize our healthcare system by developing personalized and customized medicine. The high speed genomic sequencing coupled with sophisticated informatics technology will allow a doctor in a clinic to quickly sequence a patient's genome and easily detect potential harmful mutations and to engage in early diagnosis and effective treatment of diseases. Bioinformatics tools are being used in agriculture as well. Plant genome databases and gene expression profile analyses have played an important role in the development of new crop varieties that have higher productivity and more resistance to disease.

LIMITATIONS

Having recognized the power of bioinformatics, it is also important to realize its limitations and avoid over-reliance on and over-expectation of bioinformatics output. In fact, bioinformatics has a number of inherent limitations. In many ways, the role of bioinformatics in genomics and molecular biology research can be likened to the role of intelligence gathering in battlefields. Intelligence is clearly very important in leading to victory in a battlefield. Fighting a battle without intelligence is inefficient and dangerous. Having superior information and correct intelligence helps to identify the enemy's weaknesses and reveal the enemy's strategy and intentions. The gathered information can then be used in directing the forces to engage the enemy and win the battle. However, completely relying on intelligence can also be dangerous if the intelligence is of limited accuracy. Overreliance on poor-quality intelligence can yield costly mistakes if not complete failures.

It is no stretch in analogy that fighting diseases or other biological problems using bioinformatics is like fighting battles with intelligence. Bioinformatics and experimental biology are independent, but complementary, activities. Bioinformatics depends on experimental science to produce raw data for analysis. It, in turn, provides useful interpretation of experimental data and important leads for further experimental research. Bioinformatics predictions are not formal proofs of any concepts. They do not replace the traditional experimental research methods of actually testing hypotheses. In addition, the quality of bioinformatics predictions depends on the quality of data and the sophistication of the algorithms being used. Sequence data from high throughput analysis often contain errors. If the sequences are wrong or annotations incorrect, the results from the downstream analysis are misleading as well. That is why it is so important to maintain a realistic perspective of the role of bioinformatics.

Bioinformatics is by no means a mature field. Most algorithms lack the capability and sophistication to truly reflect reality. They often make incorrect predictions that make no sense when placed in a biological context. Errors in sequence alignment, for example, can affect the outcome of structural or phylogenetic analysis. The outcome of computation also depends on the computing power available. Many accurate but exhaustive algorithms cannot be used because of the slow rate of computation. Instead, less accurate but faster algorithms have to be used. This is a necessary trade-off between accuracy and computational feasibility. Therefore, it is important to keep in mind the potential for errors produced by bioinformatics programs. Caution should always be exercised when interpreting prediction results. It is a good practice to use multiple programs, if they are available, and perform multiple evaluations. A more accurate prediction can often be obtained if one draws a consensus by comparing results from different algorithms.

NEW THEMES

Despite the pitfalls, there is no doubt that bioinformatics is a field that holds great potential for revolutionizing biological research in the coming decades. Currently, the field is undergoing major expansion. In addition to providing more reliable and more rigorous computational tools for sequence, structural, and functional analysis, the major challenge for future bioinformatics development is to develop tools for elucidation of the functions and interactions of all gene products in a cell. This presents a tremendous challenge because it requires integration of disparate fields of biological knowledge and a variety of complex mathematical and statistical tools. To gain a deeper understanding of cellular functions, mathematical models are needed to simulate a wide variety of intracellular reactions and interactions at the whole cell level. This molecular simulation of all the cellular processes is termed *systems biology*. Achieving this goal will represent a major leap toward fully understanding a living system. That is why the system-level simulation and integration are considered the future of bioinformatics. Modeling such complex networks and making predictions about their behavior present tremendous challenges and opportunities for bioinformaticians. The ultimate goal of this endeavor is to transform biology from a qualitative science to a quantitative and predictive science. This is truly an exciting time for bioinformatics.

FURTHER READING

Attwood, T. K., and Miller, C. J. 2002. Progress in bioinformatics and the importance of being earnest. *Biotechnol. Annu. Rev*. 8:1–54.

Golding, G. B. 2003. DNA and the revolution of molecular evolution, computational biology, and bioinformatics. *Genome* 46:930–5.

Goodman, N. 2002. Biological data becomes computer literature: New advances in bioinformatics. *Curr. Opin. Biotechnol.* 13:68–71.

Hagen. J. B. 2000. The origin of bioinformatics. *Nat. Rev. Genetics* 1:231–6.

Kanehisa, M., and Bork, P. 2003. Bioinformatics in the post-sequence era. *Nat. Genet.* 33 Suppl:305–10.

Kim, J. H. 2002. Bioinformatics and genomic medicine. *Genet. Med.* 4 Suppl:62S–5S.

Luscombe, N. M., Greenbaum, D., and Gerstein, M. 2001. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med*. 40:346–58.

Ouzounis, C. A., and Valencia, A. 2003. Early bioinformatics: The birth of a discipline – A personal view. *Bioinformatics* 19:2176–90.

CHAPTER TWO

Introduction to Biological Databases

One of the hallmarks of modern genomic research is the generation of enormous amounts of raw sequence data. As the volume of genomic data grows, sophisticated computational methodologies are required to manage the data deluge. Thus, the very first challenge in the genomics era is to store and handle the staggering volume of information through the establishment and use of computer databases. The development of databases to handle the vast amount of molecular biological data is thus a fundamental task of bioinformatics. This chapter introduces some basic concepts related to databases, in particular, the types, designs, and architectures of biological databases. Emphasis is on retrieving data from the main biological databases such as GenBank.

WHAT IS A DATABASE?

A *database* is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria. Databases are composed of computer hardware and software for data management. The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information. Each record, also called an *entry*, should contain a number of fields that hold the actual data items, for example, fields for names, phone numbers, addresses, dates. To retrieve a particular record from the database, a user can specify a particular piece of information, called *value*, to be found in a particular field and expect the computer to retrieve the whole data record. This process is called *making a query*.

Although data retrieval is the main purpose of all databases, biological databases often have a higher level of requirement, known as *knowledge discovery*, which refers to the identification of connections between pieces of information that were not known when the information was first entered. For example, databases containing raw sequence information can perform extra computational tasks to identify sequence homology or conserved motifs. These features facilitate the discovery of new biological insights from raw data.

TYPES OF DATABASES

Originally, databases all used a flat file format, which is a long text file that contains many entries separated by a *delimiter*, a special character such as a vertical bar (|). Within each entry are a number of fields separated by tabs or commas. Except for the

raw values in each field, the entire text file does not contain any hidden instructions for computers to search for specific information or to create reports based on certain fields from each record. The text file can be considered a single table. Thus, to search a flat file for a particular piece of information, a computer has to read through the entire file, an obviously inefficient process. This is manageable for a small database, but as database size increases or data types become more complex, this database style can become very difficult for information retrieval. Indeed, searches through such files often cause crashes of the entire computer system because of the memory-intensive nature of the operation.

To facilitate the access and retrieval of data, sophisticated computer software programs for organizing, searching, and accessing data have been developed. They are called *database management systems.* These systems contain not only raw data records but also operational instructions to help identify hidden connections among data records. The purpose of establishing a data structure is for easy execution of the searches and to combine different records to form final search reports. Depending on the types of data structures,these database management systems can be classified into two types: *relational database management systems* and *object-oriented database management systems*. Consequently, databases employing these management systems are known as *relational databases* or *object-oriented databases*, respectively.

Relational Databases

Instead of using a single table as in a flat file database, relational databases use a set of tables to organize data. Each table, also called a *relation*, is made up of columns and rows. Columns represent individual fields. Rows represent values in the fields of records. The columns in a table are indexed according to a common feature called an *attribute*, so they can be cross-referenced in other tables. To execute a query in a relational database, the system selects linked data items from different tables and combines the information into one report. Therefore, specific information can be found more quickly from a relational database than from a flat file database.

Relational databases can be created using a special programming language called *structured query language* (SQL). The creation of this type of databases can take a great deal of planning during the design phase. After creation of the original database, a new data category can be easily added without requiring all existing tables to be modified. The subsequent database searching and data gathering for reports are relatively straightforward.

Here is a simple example of student course information expressed in a flat file which contains records of five students from four different states, each taking a different course (Fig. 2.1). Each data record, separated by a vertical bar, contains four fields describing the name, state, course number and title. A relational database is also created to store the same information, in which the data are structured as a number of tables. Figure 2.1 shows how the relational database works. In each table, data that fit a particular criterion are grouped together. Different tables can be linked by common data categories, which facilitate finding of specific information.