CHAPTER

4

The Beginning of Bioinformatics*

OUTLINE

4.1 MARGARET DAYHOFF, RICHARD ECK, ROBERT LEDLEY, AND THE BEGINNING OF BIOINFOMATICS

Although bioinformatics is one of the buzzwords in the post-genomic era, it is by no means a completely new discipline. The beginning of the pioneering work by Margaret Dayhoff, Richard Eck, and Robert Ledley in computer-aided analysis of protein data goes back to the period around 1960. Dayhoff, Eck, and Ledley capitalized on their experience and training in computing, mathematics, and life sciences in collecting and organizing protein sequences, sequence analysis, and studies of protein evolution.[1,2,3] Their work could be regarded as the direct ancestor of modern bioinformatics. In 1965, Dayhoff, Eck, and a couple of colleagues compiled the first **Atlas of Protein Sequence and Structure**, which had ~50 sequences known at the time. The second volume was published in 1966 and had a little over 100 sequences. This compilation of protein sequence and structure information was the predecessor of the current gene and protein databases that form the backbone of contemporary bioinformatics. In subsequent years, as more and more protein sequences were reported, the Atlas grew in size and popularity under the leadership of Dayhoff. Eventually, this database became The **Protein Information Resource (PIR) database**, now maintained at Georgetown University.

Margaret Dayhoff was a professor at Georgetown University Medical Center. As an independent researcher, Dayhoff brought her background of mathematics, chemistry, and computing to address problems in biology, particularly protein chemistry, and became the pioneer in the application of mathematics and computational methods to biochemistry. One of her most important contributions was developing, together with Richard Eck, the single-letter code for amino acids that is used by all protein analysis tools. She developed a computer algorithm for protein-sequence alignment, which was (correctly) thought to reveal their evolutionary history.

Richard Eck studied chemical engineering and plant biology. In 1961, Eck published a paper in *Nature* in which he compared all the sequences of hemoglobin variants, and other proteins such as insulin, from different species. He realized that the information on amino-acid sequences could be organized in different ways in order to produce specific patterns. He also identified numerous amino-acid substitutions in proteins and noted that the pattern of substitutions was not random. In a conference in 1964, Eck presented a **cryptogrammic** method to trace the evolution of proteins. He suggested that, using this result, one could

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

calculate the degree of relatedness of each protein with reference to its ancestors, and draw a family tree in which the distances between the branches represented a quantitative measure of relatedness. Thus, Eck outlined the basis of reconstruction of a phylogenetic tree.

Robert Ledley, who studied theoretical physics and dentistry, envisioned an important application of computers to sequence analysis. He suggested that after the polypeptide chain is cut into many overlapping fragments, whose sequences could be determined by peptide sequencing, the fragment reassembly of partial sequences to obtain full sequences could be done using computers. Thus, Ledley suggested that computers could assist biochemists in their efforts to determine protein sequences. He invited Dayhoff to join the staff of National Bureau of Standards (NBRF; later the National Institute of Standards and Technology, or NIST) in 1960 to continue investigating this question. Dayhoff and Ledley wrote FORTRAN programs that could direct the assembly of partial peptide sequences in the right order in less than 5 minutes.

Both Dayhoff and Eck became involved in evolutionary studies of proteins while Ledley continued with his interest in the application of computers in biology. Dayhoff started playing an increasingly important role in protein-sequence analysis and continued to contribute to evolutionary biology based on her studies on protein sequences. She published the first reconstruction of a phylogenetic tree using a maximum parsimony method, discussed in Chapter 9. She also developed the first amino-acid substitution matrix for studying protein evolution, called the **PAM matrix**. PAM stands for **point accepted mutation** (also referred to as **percent accepted mutation**) because it represents accepted point mutation per 100 amino acid residues. A publication by Dayhoff in the popular science journal *The Scientific American*, entitled *Computer Analysis of Protein Evolution*,[4] can be regarded as one of the most important initial publications in bioinformatics and molecular phylogenetics. For her enormous pioneering contributions, Margaret Dayhoff is popularly regarded as the founder of modern bioinformatics.

4.2 DEFINITION OF BIOINFORMATICS

The term "bioinformatics" was coined by Paulien Hogeweg and Ben Hesper in 1978.[5] In a recent review article recapitulating the history of bioinformatics, Hogeweg stated that the term had been used by Hogeweg and Hesper since the beginning of the 1970s, but was formally coined in 1978 in an article written in Dutch. In the beginning, the term was used to mean the study of informatic processes in biotic systems. Bioinformatics is basically informatics as applied to biology—that is, computer-aided analysis of biological data. There are many definitions/descriptions of bioinformatics; some of these definitions make no distinction between bioinformatics and computational biology as a whole. Luscombe et al.[6] defined bioinformatics as follows:

> Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the information associated with these molecules, on a large-scale.

Higgs and Attwood[7] provided two definitions of bioinformatics that are same in spirit but stated in two different ways:

> (1) Bioinformatics is the development of computational methods for studying the structure, function, and evolution of genes, proteins and whole genomes; and (2) bioinformatics is the development of methods for the management and analysis of biological information arising from genomics and high-throughput experiments.

Therefore, for molecular biologists, bioinformatics is the discipline of computer-aided analysis of information relating to genes, genomes, and their products. In other words, for all practical purposes, bioinformatics can be regarded as **computational molecular biology**, that uses computational techniques to study the structure, function, regulation, and interactive network of genes and proteins. The ultimate goal is to analyze and predict the structure, organization, function, regulation, and dynamics of the entire genome of an organism.

4.3 BIOINFORMATICS VERSUS COMPUTATIONAL BIOLOGY

Computational biology is an umbrella term that includes any subdiscipline in biology that uses computer-aided analysis, modeling, and prediction. Some examples include the modeling of predator–prey relationships in an ecosystem, the modeling and prediction of population and community dynamics in an ecosystem, quantitative structure–activity analysis and prediction of the biological effects of chemicals, prediction of metabolic fate of chemicals in vivo, and pharmacokinetic modeling of drugs and xenobiotics, etc. In contrast, bioinformatics can be regarded as computational molecular biology, as indicated above. Therefore, according to the definitions discussed in this book, computational biology is much broader in scope and bioinformatics is a part of it. Bioinformatics, like other areas of computational biology, is essentially a multidisciplinary science because it uses techniques and concepts from a number of disciplines, such as molecular biology and biochemistry, computer science, statistics and mathematics, and informatics (information science).

4.4 GOALS OF BIOINFORMATIC ANALYSIS

The ultimate goal of bioinformatics is to be able to predict the biological processes in health and disease. In order to acquire such an ability, a thorough understanding of the biological processes is necessary. Therefore, the proximate goal of bioinformatics is to develop such an understanding through analysis and integration of the information obtained on genes and proteins, as well as to develop new tools and continuously improve the existing set of tools for diverse types of analyses. Bioinformatics also aims to develop tools that help in the management of and access to data and information, including improved search and retrieval capability of genomic data and information from various types of databases. Some examples of common bioinformatic tools and analyses that are continuously being improved and refined are: data capture and storage capability; the usability of databases; data analysis; nucleic acid and protein sequence analysis and sequence annotation; structural analysis of proteins and prediction of protein structure, including three-dimensional (3D) structure; protein domain prediction; gene prediction; analysis of functional studies; analysis of gene and protein networks; and phylogenetic analysis.

The analytical tools in bioinformatics are computer algorithms and statistics. Improvements in the capacity of existing tools and the development of new tools are both driven by the need for newer dimensions and greater speed of analysis, as well as the ability to handle an ever-increasing amount of data. However, the success and prediction accuracy of bioinformatic analysis ultimately depends on our knowledge of the biology of organisms. Therefore, as more data accumulate in the databases and more scientific information becomes available, the progress of science and its prognostic ability will require and hence dictate the development of new bioinformatic tools. Acquisition of more data and information, storage of all that information, expansion of databases, new strategies needed for analysis, and advances in computing power are all expected to facilitate the analysis of large volumes of data and discovery of new biological principles and insights from which unifying principles of life and its evolution can be discerned.

4.5 BIOINFORMATICS TECHNICAL TOOLBOX

Bioinformatic analysis requires data (such as sequence information), databases, and analysis tools. Databases are built from data obtained through wet laboratory experiments. Some of the original nucleotide- and protein-sequence databases were created more than 30 years ago. Subsequently, information from these original databases was utilized to create curated and more refined databases to meet specific research needs. With the advances in genomics, proteomics, and metabolomics, particularly with the development of disciplines like pharmacogenomics and toxicogenomics, the need for storage of and access to the newly created datasets has led to the development of further specialized databases. Through the collaboration of academic, corporate, and regulatory scientists, standards have been developed as to how to submit a specific type of data to the relevant databases. A more detailed discussion of various databases will be undertaken in Chapter 5.

The bioinformatics technical toolbox provides analysis tools (algorithms) and visualization techniques of the data generated through high-throughput experiments, such as high-throughput sequencing, microarray analysis, mass spectrometry, and other proteomic techniques. The analysis tools are computer based (software), and the development of newer tools is driven by various needs, such as an increased need for handling the huge body of data, faster analysis, expanded scope of the analysis, multiple simultaneous analyses, to name a few. A few examples of software-driven analysis that have tremendously facilitated bioinformatics research are:

Analysis of nucleotide sequences
Detection of single nucleotide polymorphisms (SNPs) and copy number variation (CNV)
Understanding the sequence features and differences between coding and noncoding regions
Alignment of nucleotide sequences
Prediction of open reading frames (ORFs), restriction-enzyme cutting sites in DNA, various *cis*-acting regulatory DNA elements in the gene, and putative miRNA-encoding sequences in the genome
Gene-expression analysis
Designing probes and primers
Analysis of protein sequences
Alignment of amino-acid sequences
Prediction of protein structure (including 3D structure), protein–protein interactions, post-translational modifications of proteins, hydrophilicity/hydrophobicity and potential

antigenicity of proteins, and various protein domains, such as transmembrane domains

Prediction of phylogenetic relationships among proteins.

In addition, gene-expression analysis information has led to the development of systems biology tools that can perform simulation, steady-state analysis, network identification, complex behavior analysis of the system, and various other tasks.

References

1. Strasser BJ. *J Hist Biol* 2010;**43**:623−60.
2. Lee J. *Prot Sci* 2007;**16**:1509−10.
3. Doolittle RF. *PLoS Comput Biol* 2010;**6**:e1000875.
4. Dayhoff MO. *Sci Am* 1969;**221**:86−95.
5. Hogeweg P. *PloS Comput Biol* 2011;**7**:e1002021.
6. Luscombe NM, et al. *Methods Inf Med* 2001;**40**:346−58.
7. Higgs PG, Attwood TK. *Bioinformatics and molecular evolution*. MA: Blackwell; 2005.