1

Introduction

1.1 The Genesis of Bioinformatics

Bioinformatics is a discipline which originally arose for the utilitarian purpose of introducing order into the massive data sets produced by the new technologies of molecular biology. These techniques originated with large-scale DNA sequencing and the need for tools for sequence assembly and for sequence annotation, i.e., determination of locations of protein-coding regions in DNA. A parallel development was the construction of sequence repositories. The crowning achievement has been the sequencing of the human genome and, subsequently of many other genomes.

Another new technology, which has started to provide wealth of new data, is the measurement of multiple gene expression. It employs various physical media, including glass slides, nylon membranes, and other media. The idea is to expose a probe (a DNA chip) including thousands of DNA nucleotide sequences, each uniquely identifying a gene, to a sample of coding DNA extracted from a specimen of interest. Multiple-gene-expression techniques are usually employed to identify subsets of genes discriminating between two or more biological conditions (supervised classification), or to identify clusters in the gene sample space, which leads to a classification of both samples and genes (unsupervised classification). Analysis of gene expression data has led to new developments in computational algorithms: existing computational techniques, with their origin in computer science, such as self-organizing maps and support vector machines, and of statistical origin such as principal-component analysis and analysis of variance, have been adapted, and new techniques have been developed.

The next step in the development of the technology includes proteomic techniques, which allow measurements of the abundance and activity of thousands of protein species at once. These are usually multistep procedures. The initial phase involves physical separation of proteins from the sample according to one or more (typically two) variables, for example molecular weight and isoelectric point. This is physically accomplished using two-dimensional gels, on which different proteins can be spotted as individual clusters. The next step involves identification of proteins sampled from different spots on the gel. This involves cleavage of amino acid chains and producing mass spectra using extremely precise mass spectrometry machines. Finally, on the basis of the distribution of molecular weights of the fragmented chains, it is possible to identify known proteins or even to sequence unknown ones. Various more refined versions of the technology exist, which allow the labeling of activated proteins, various protein subsets, and so forth.

The interpretation of proteomic data has led to the development of warping and deconvolution techniques. Two-dimensional protein gels are distorted with respect to the perfect Cartesian coordinates of the two variables describing each protein. To allow comparison with standards and with results obtained under other experimental conditions, it is necessary to transform the gel coordinates into Cartesian ones, a procedure known as warping. As mentioned above, after this is accomplished, we may analyze a gel spot representing a protein, using mass spectrometry. Deciphering the sequence of the polypeptide chain using mass spectrometry of fragments $5 - 10$ amino acids long is accomplished using deconvolution.

One of the more notable consequences of the developments in genomics and proteomics has been an explosion in the methodology of genetic and metabolic networks. As is known, the expression of genes is regulated by proteins, which are activated by cascades of reactions involving interactions with other proteins, as well as the promotion or inhibition of the expression of other genes. The resulting feedback loops are largely unknown. They can be identified by perturbing the system in various ways and synthesizing a network on the basis of genomic and proteomic measurements in the presence of perturbations. A variety of network types can be used, varying from Boolean networks (discrete automata) and probabilistic versions of them, to Bayesian networks and others. Although these techniques are still unsatisfactory in practice, in many cases they have allowed us to gain insight into the structure of the feedback loops, which than can be analyzed using more conventional tools, including, for example, systems of nonlinear differential equations.

1.2 Bioinformatics Versus Other Disciplines

Bioinformatics has been developed in the space, which was already occupied by a number of related disciplines. These include quantitative sciences such as

- mathematical and computational biology,
- biometry and biostatistics,
- computer science,
- cybernetics,

as well as biological sciences such as

- molecular evolution,
- genomics and proteomics,
- genetics, and
- molecular and cell biology.

It might be argued that bioinformatics is a direct extension of mathematical and computational biology into the realm of new, massive data sets. However, the sheer size of this mass of data creates qualitatively new situations. For example, any serious query of the data requires writing computer code and/or placing the data within a suitable database. The complexity of the databases varies enormously, reaching the highest proportions in databases designed to handle information about metabolic pathways. Even determining what should be the subject of a query involves computer-intensive methods.

As an example, let us consider the problem of finding enough homologous DNA sequences to carry out an evolutionary analysis of homologous proteins coded by these sequences in different organisms. To accomplish this, one has to use a set of computerized tools, known as BLAST, which has the ability to search for sequences above a certain level of similarity and to assign statistical similarity scores to potential homologs. The probabilistic theory of BLAST involves considerations of how unlikely it is for two sequences of given length to display a given level of similarity.

Another interesting example concerns carrying out statistical comparisons between gene expression levels obtained using DNA microarrays. Here, we have to deal with comparisons of a limited number of microarrays, each yielding a data vector of high dimension. This is a situation which is exactly opposite to the usual statistical paradigm, according to which a large sample of low-dimensional data is considered most useful. Even worse, comparisons are frequently carried out gene-by-gene, leading to potential repeated-testing problems. This problem becomes even more serious when we realize that large subsets of genes may have correlated expressions. Under such circumstances, the only statistical tools which make it possible to determine whether differences found are significant, are permutation tests. These latter are often computationally intensive,

A major issue in bioinformatics is the combinatorial complexity of algorithms, which can be insurmountable. An example stemming from the field of molecular evolution is the construction of phylogenetic trees of sequences using the maximum-likelihood method. The space of trees with more than 10 nodes is so enormous that there is no way an exhaustive search might be carried out. Instead, various heuristics and suboptimal searches are used. This is an important point, since, as noted later, evolutionary changes of biological sequences can be treated as a result of an experiment not requiring a new laboratory. This is discussed later in the context of identification of active sites of proteins.

Another example of a typically bioinformatic problem is provided by polymorphisms in the human genome. As is known, any two human DNA sequences differ at random points located, on average several hundred nucleotides apart. These are the single-nucleotide polymorphisms (SNPs). Therefore, there exists an enormous multitude of sequence variants. At the same time, the human genome sequence is based on only a few individuals. This illustrates the difference with respect to classical human genetics, which attempts to elucidate the role of genetic variability at a limited number of loci at a time. With the onset of mass sequencing of either entire genomes or major portions of genomes, analysis of their genetic and evolutionary relationships will require increased computational power and new data structures.

1.3 Further Developments: from Linear Information to Multidimensional Structure Organization.

Many widely used methods of bioinformatics hinge upon the linear structure of genomic information. This includes sequencing and annotation, but also sequence retrieval and comparison. A natural toolbox for problems of this nature is provided by hidden Markov models (HMMs) and the Viterbi algorithm based on dynamic programming. The idea of the Viterbi algorithm is to find the most likely estimate of the Markov process underlying a given biological process, based on the so-called emissions, i.e., the limited available observations of the process. The solution is obtained recursively, following the dynamic programming paradigm. A typical application of the Viterbi algorithm arises when the Markov process describes some feature of the genetic/genomic information distributed along the DNA sequence (this can be some functionality of the DNA region) and the emissions are constituted by the sequence of DNA nucleotides. An example is the identification of promoter regions of genes. However, the Viterbi algorithm can be defined for Markov processes evolving on very general spaces. For example, consider the space of nested quasi-palindromic motifs, which is equivalent to all possible secondary structures of RNA molecules, endowed with a Markov process defined as a stochastic algebra of admissible rules by which the motifs can be created. This framework makes possible to define a Viterbi algorithm for identification of the structure, based on the sequence. Other interesting applications of the Viterbi algorithm arise when we attempt to build phylogenetic trees of sequences involving a variable substitution rate along the sequence. This extension to branching structures is the foundation of the Felsentein–Churchill algorithm for maximum likelihood trees, discussed later.

Biological information is translated into the structure and function of biomolecules, which in turn form higher-level structures. The simplest example is protein folding. Proteins are active because of their spatial conformation and the occurrence of active centers, which interact with other molecules. Quantitative studies of these features can be accomplished in various ways. A direct approach involves computations of protein folding based on energy functions. Again, dynamic programming can be used to reduce the computational burden. If this is accomplished, or if the structure is known from X-ray crystallography, it is possible to consider computations of active centers of proteins based on the geometry of their surfaces. The interaction of proteins may be approached computationally by solving the docking problem, employing methods of computational geometry similar to those used in robotics. These and related computations are involved and time-and memory-consuming.

An alternative approach is based on the notion of evolution as a laboratory. By following the evolution of biomolecules, it is possible to infer their function and the relationships between them. Example of this approach is the evolutionary trace method of Lichtarge. In this method, homologous amino acid sequences from a number of species are used to infer a phylogeny. Subsequently, this phylogeny forms a basis for classification of the amino acids in the sequence, based on their conservation in branches of the tree of different order. The amino acids which are conserved best are likely to belong to the active center. This method has led to confirmed predictions of the active sites. Similarly, the Felsentein–Churchill algorithm mentioned above allows identification of amino acids, which have evolved slowly. These will be candidates for belonging to the active center.

The new branches of bioinformatics will require the creation of new databases and continued work on purely informatic structures such as ontologies, which allow retrieval of information with a very rich structure.

1.4 Mathematical and Computational Methods

At present, virtually all branches of science use mathematical methods as parts of their research tools. Science has entered a phase of mathematics invading other disciplines. This is because the concepts in all areas of science are becoming more and more mature and precise, and mathematical tools are flexible and generalizable.

Without exaggeration, we can say that the majority of the methods of applied mathematics are used as tools in bioinformatics. So, is there anything peculiar about using mathematical modeling in bioinformatics? Among the tools of applied mathematics some are of special importance, namely probability theory and statistics and algorithms in computer science. A large amount of research in bioinformatics uses and combines methods from these two areas. Computer-science algorithms form the technical background for bioinformatics, in the sense that the operation and maintenance of bioinformatic databases require the most up-to-date algorithmic tools. Probability and statistics, besides being a tool for research, also provides a language for formulating results in bioinformatics.

Other mathematical and computational tools, such as optimization techniques with dynamic programming, discrete-mathematics algorithms, and pattern analysis methods, are also of basic importance in ordering bioinformatic data and in modeling biological mechanisms at various levels.

The first part of the book, on mathematical and computational methods is intended to cover the tools used in the book. The presentations of methods in this part are oriented towards their applications in bioinformatics. In the second part of this book, practical uses of these methods are illustrated on the basis of the rather large number of research papers devoted to the analysis of bioinformatic data. Sometimes some further developments of methods are presented, together with the problem they apply to, or some references are given to the derivation of the algorithm. Description of applied mathematical methods is organized into several sections corresponding to logical grouping of methods.

Our presentation of the mathematical approaches is rather descriptive. When discussing mathematical methods we appeal to comprehension and intuitive understanding, to their relations to bioinformatic problems and to cross-applications between items we discuss. This approach allows us to go through a variety of methods and, hopefully, to sketch a picture of bioinformatics. Despite avoiding much of the mathematical formalism we have tried to keep the presentation sufficiently clear and precise. All chapters are accompanied by exercises and problems, which are intended to support understanding of the material and often show further developments. Their levels of difficulty varies, but generally they are rather non trivial.

1.4.1 Why Mathematical Modeling?

What is mathematical modeling? By mathematical modeling, we understand describing and reflecting reality by using formalized tools. Models can be of very different types: stochastic or deterministic, descriptive or mechanistic, dynamic or static. Mathematical models can pertain to phenomena in many different areas, for instance physics, chemistry, biology, engineering, or economics.

How do we develop models? Models are developed by combining, comparing, or verifying hypotheses versus empirical observations. We develop models by using the laws of nature, physics, chemistry, and biology. We apply principles of conservation and/or variational extremum principles, which lead to balances and to differential or difference equations for the evolution of the state of a system. Models can include discrete events and random phenomena.

What is the benefit of using mathematical models? Using mathematical models allows us to achieve a better understanding and to organize better our knowledge about the underlying mechanisms and phenomena. Sometimes models can change qualitative understanding to quantitative knowledge. Models can allow us to predict future events from present observations. Models can be helpful in programming and planning our control and design actions.

What is specific in modeling in biology and molecular biology? Compared with models in physics and classical chemistry, models in (molecular) biology pertain to more complex phenomena. Following from this there is usually a greater extent of simplification that needs to be applied when building the model. The large individual variation leads to a substantial element of randomness, which needs to be incorporated into the model.

1.4.2 Fitting Models to Data

An element which is present in all models is simplifying hypotheses. The benefit in using a mathematical model is often related to solving the compromise between the extent of simplification in the model and the precision in predicting data. Complicated models are usually less reliable and less comprehensive. Oversimplified models can ignore important phenomena.

The research work that forms part of modeling involves model building or model learning, applying the model to the data and model modification. After enough experience has been gained by repeated application of these elements of modeling research, models often start bringing benefits.

One crucial element is verifying a model versus the data, which very often starts from fitting free parameters of the model. This involves tasks such as identification and parameter estimation, solved by various methods of static, dynamic or stochastic optimization. Among optimization methods the least squares method deserves special attention owing to its reliability and very vast range of application.

If one assumes model with many free parameters, one has substantial flexibility in fitting the model to the data. The extreme case is called “black box modeling”, which means fitting the parameters of standardized models to the measurements without inquiring about the nature of the underlying processes and phenomena.

1.4.3 Computer Software

Both fitting to data and analyzing the predictions of mathematical models is done by using computers with appropriate software. There is a variety of computer software environments for all platforms, and choosing the appropriate program for the computational aspects of the research being done is an important issue. Some very useful programming environments are the high-level programming languages for supporting engineering and scientific computations Matlab, Mathematica, Maple, R. Several computational examples in this book were programmed using Matlab. Matlab can be equipped with toolboxes, which include many of the algorithms described in this book. For some specialized tasks one may need programming languages of lower level, such as C, C++, Delphi, Java.

We should also mention the numerous Internet servers offering specialized computations in the field of bioinformatics, such as aligning sequences against databases, predicting 2D and 3D structures of proteins and RNA and so forth. Some of them are mentioned or discussed later in this book.

1.5 Applications

Facts in biology and biochemistry become established when they are seen in a biological or a biochemical experiment or, better, in several independent experiments. Knowledge develops in biology and biochemistry in this way. There are two aspects of the development of biology and biochemistry concerning its relation to bioinformatics. First, with the development of experimental techniques, the number of findings and discoveries in biology and biochemistry has become so large, that efficient access to the information requires the use of the most advanced informatic tools. Second, browsing and analyzing data in bioinformatic databases allows or helps us to predict facts in biology and biochemistry or to propose new hypotheses. These hypotheses can be used for designing new experiments. There are several well-established paths in which bioinformatics can be applied in this second way. After the genome of a new organism has been sequenced, then by using knowledge about the structure and organization of genomes and the contents of genomic databases, researchers can find the genes and compare them with their homologs in other organisms. Inside genes, coding sequences can be identified, leading to amino acid sequences of proteins. These approaches can be used in a variety of types of research. Information obtained from comparing genomes can be used for inferring the ancestry of organisms and also for predicting the functions of genes and proteins. Comparing sequences of amino acids in proteins in different organisms allows one to infer their functionally important sites and active sites. By combining computational methods with browsing protein databases, one can improve the methods for drug design. For example, when the sequence of a virus causing a disease has been found then it is often searched for regions coding for proteins. Next, using the hypothesis that these proteins are important in the activity of the virus in the human organism, design of the appropriate treatment can focus on drugs blocking their activity.

Bioinformatic databases contain massive amounts of experimental data. Browsing and analyzing these data is fascinating and will surely lead to many interesting discoveries. The developing projects concerned with searching for interesting information in bioinformatic databases belong to the most vital area in scientific research.

It is important to stress here the interdisciplinary aspects of the research in bioinformatics. A search through bioinformatic databases is often initiated by posing a question related to some biological problem. The bioinformatic project then involves designing the computational and algorithmic aspects of the search or browsing. The results are most valuable when they lead to answering the question, to improved understanding or to interesting biological interpretations.

In the second part of this book, we have organized the material such that the biological and biochemical aspects are treated with enough care to explain the motivation for pursuing research in bioinformatics. The second part of the book includes seven chapters, each devoted to a specific area. We start with two chapters, on sequence alignment and molecular phylogenetics, devoted to specific methodologies applicable in many contexts, which are discussed later. In the chapter on sequence alignment, we present the methodologies and their relation to optimization and to computer-science algorithms. In the chapter on molecular phylogenetics we discussed basic approaches of reconstructing phylogenetic trees, using appropriate tools of optimization and statistics. We also included a section on coalescence, which (i) allows us to understand the processes behind the formation of phylogenetic trees, and (ii) illustrates some new applications of phylogenetics, such as inferring demographic scenarios from molecular data. The next three chapters are devoted to biological items, namely genomics, proteomics and RNA. These chapters include, in their introductory parts, the basics of the underlying biological and biochemical facts. Next, mathematical modeling methods and their relations to experimental approaches are presented. The chapter on DNA microarrays is focused on the biological process of gene expression and the associated technology of biological assays, as well as related mathematical and computational approaches. Owing to its importance and the large number of research papers and monographs in the field, it deserves special attention. We have provided a description of DNA microarray technology in the introductory part. Then we discuss mathematical modeling in the context of analyzing gene expression profiles. Finally, the last chapter is devoted to bioinformatic databases and other bioinformatic Web sites and services. In this short chapter, we have aimed to give an overview of some of the internet resources related to bioinformatics.

Most of the chapters have a set of exercises at the end. Some exercises are problems aimed at supporting understanding of presented ideas and often completing or adding some elements of derivations of methods. Other exercises are projects, which often involve issues such as developing computer programs and studying their application to solving problems. Many of the projects suggest downloading publicly available software and/or using some of internet bioinformatic depositories on the Internet. In these projects we have suggested many possibilities, which we are fairly sure will help to develop our understanding of some problems and may lead to interesting results.