Bioinformatics

Sequence and Genome Analysis

生物信息学

——核苷酸链与基因组分析

SECOND EDITION

David W.Mount

世界图书出版公司

Bioinformatics
Sequence and Genome Analysis
生物信息学
——核苷酸链与基因组分析

Edited by

David W. Mount

CSHL PRESS

世界图书出版公司

西安 北京 广州 上海

COLD SPRING HARBOR LABORATORY PRESS

陕版出图字:25-2004-130
图书在版编目(CIP)数据

生物信息学——核苷酸链与基因组分析/(美)大卫(David W.Mount)主编.—西安:世界图书出版西安公司,2006.7
ISBN 7-5062-4764-X

Ⅰ.生... Ⅱ.大... Ⅲ.生物信息学工程—英文
Ⅳ.R517

中国版本图书馆 CIP 数据核字(2005)第 081473 号

生物信息学
——核苷酸链与基因组分析

主　　编　(美)David W. Mount
策　　划　任卫军
责任编辑　汪信武　段　晖
封面设计　高宏超

出版发行　世界图书出版西安公司
地　　址　西安市北大街 85 号
邮　　编　710003
电　　话　029-87285225　87285507(医学读者俱乐部)　87214941(市场营销部)
　　87235105(总编室)
传　　真　029-87279075　87279676
经　　销　各地新华书店
印　　刷　人民日报社西安印务中心
开　　本　889 mm×1194 mm　1/16
印　　张　44
字　　数　1400 千字
彩　　插　16 页

版　　次　2006 年 7 月第 1 版　2006 年 7 月第 1 次印刷
书　　号　ISBN 7-5062-4764-X/R·517
定　　价　368.00 元

Preface to the Second Edition

This second edition of *Bioinformatics: Sequence and Genome Analysis* has a wider target audience than the first edition and is designed both for biologists who want to learn about computational and statistical methods and for computational scientists who want to learn about biology, especially genetics and genomics. Chapter guides introduce the basic computational/statistical and biological backgrounds needed for each chapter. Additional new chapter elements include a list of topics that should be learned in the chapter, Web search terms at the end of each chapter (URLs of stable Web sites are still shown in tables and main text), and problem sets at the end of each chapter that emphasize chapter concepts and skills. Finally, all supplementary material on the Web site, mostly for Chapter 3, has now been summarized in the text so that the material is all in one place. All of the original chapters have been updated, revised, and rewritten.

Three new chapters have been added to this second edition. Chapter 4, which covers probability and statistical analysis of sequence alignments and was part of Chapter 3 in the first edition, now includes additional background about hypothesis testing and testing the accuracy of predictions based on sequence analysis. Chapters 12 and 13 explore subjects that were not covered in the first edition—Perl programming and analyzing microarrays. These chapters, which are given at an advanced level and assume some background on the part of the reader, were added so that all of the most relevant areas of bioinformatics could be included in a single reference text. These chapters are state of the art and constitute an invaluable addition to the second edition. The material in these chapters was contributed by colleagues at the University of Arizona, to whom I am very grateful. Going through drafts and revisions was a time-consuming and painstaking task, and all three contributors were always accommodating and generous with their time.

Chapter 12 was contributed by Nirav Merchant and Susan Miller, who are experienced computer systems and software specialists. This chapter gives considerable detail on using and generating Perl scripts and modules for a variety of tasks. Also covered are data formats and setting up relational databases. The chapter presents many sample Perl scripts that may be downloaded from the book Web site and used as templates for projects. We anticipate that Chapter 12 will be the starting point for many useful Perl programs that support large genome projects.

Chapter 13 was contributed by Dr. David Henderson, a statistician who specializes in QTL analysis and statistical design and analysis of experiments, and who has considerable experience with microarray experiments. It is designed to help biologists extract significant information from microarrays regarding gene expression. Biologists are not used to planning experiments that involve large data sets and extracting information from those experiments. Microarray experiments are also particularly troublesome for two reasons: First, there are many sources of noise in expression data, and finding which genes are showing a significant change in expression over and above this noise can be a difficult undertaking; second, the outcome of

a microarray experiment is often a long, bewildering list of genes that is difficult to sort through. We provide guides for dealing with both of these difficulties. First, guidelines for designing experiments for a specific purpose or research objective so that the noise can be removed from microarray data are given. Second, methods for finding genes that are changing significantly are described. Third, methods of analysis, including clustering methods based on different criteria and classifier-(biomarker)-finding and validation methods, are presented. Finally, the programming resources that are needed to perform these tasks are provided. The goal of Chapter 13 is, given this background information, to make it possible to design microarray experiments that yield a maximum possible amount of useful information.

Many people contributed to the second edition by providing comments on the first edition that were very helpful. At the top of the list is Yasushi Okazaki who, while translating the first edition of this book into Japanese, provided many comments and corrections. Others who provided help on numerous occasions included John Clark, Gabriel Dorado, Dan Flath, Toni Kusalik, and Etsuko Moriyama. This text would not have been possible without the support of my bioinformatics colleagues Ritu Pandey and Rob Klein, and financial support from the University of Arizona, in particular through Vicki Chandler, Gene Gerner, Rich Jorgenson, Serrine Lau, Ray Nagle, and Dan Von Hoff. I also thank Walt Klimecki for Figure 11.2 and Roger Miesfeld for help with Figure 9.13A. I am also grateful to Beth Nickerson for providing valuable critiquing and comments on many of the chapters, and to Pick Wei Lau for helping to read the galley proofs, and to Eric Shen for collecting and sorting through comments from the book Web site.

Last, but not least, I am grateful to the staff at Cold Spring Harbor Laboratory Press for their assistance. This project would not have been successful without the guidance of Judy Cuddihy, who provided an improved chapter format and exceptionally useful comments and critiques throughout the text; she kept the book to a reasonable schedule and offered needed encouragement and support. Mary Cozza guided me through the references; Patricia Barker, Kathleen Bubbeo, Daniel deBruin, and Susan Schaefer worked efficiently on production under a tight schedule; and Jan Argentine and Denise Weiss oversaw the efforts of their staffs working on this book.

D.W.M.
May 2004, Tucson

Preface to the First Edition

This book is written mainly for biologists who want to understand the methods of sequence and structure analysis. I strongly believe that a person using a computer program should understand how it works. Accordingly, one of my main objectives is to help biologists appreciate the underlying algorithms used and assumptions made, as well as limitations of the methods used and strategies for their use. To this end, I have tried to avoid complex formulas and notations and to give instead simple numerical examples whenever possible. I hope that the book will also be of interest to computational biologists who want to learn a little more about the biological questions related to the field of bioinformatics. This book is intended to be a laboratory reference text, as well as a textbook for a course in bioinformatics, rather than a user guide for a specific set of sequence analysis programs.

Most of the chapters include a flowchart that is designed to propose an orderly use of the methods that are discussed in the chapter. There are very few examples of these types of charts and they are quite difficult to produce, requiring assumptions and oversimplifications that may not always be justified. I hope that these charts will be useful for the less experienced in this field, but I expect that the more-experienced practitioners in the field will have other, probably better, ways of achieving the same goal.

There are many references to Web sites and FTP locations where these methods may be applied or programs obtained. In some cases, as for the commonly used and important BLAST and CLUSTALW programs, I have provided a great deal of information about using the program and analyzing the results. However, there are many other important tools and approaches available for biological sequence and genome analysis and I have tried to cover as many of them as possible, given time and space limitations. I paid particular attention to simpler types of sequence analyses, e.g., searching for restriction sites, translating sequences, and compositional analysis. There are many commercial and noncommercial packages for performing these tasks, and commercial packages for genome analysis are now appearing.

In writing this book, my first, I found that the amount of information available in the published literature was far more than I could include. I have tried to be thorough and to cover the most significant problems in sequence and genome analysis, but there are also many excellent papers that have not been cited for reasons of time and space, and I apologize to colleagues whose valuable contributions are not mentioned. Because of the space limitations of a printed text, and the ever-changing nature of bioinformatics, material not included in the book, as well as links to all of the Web sites cited, examples, and problems, will appear on a special Web site for the book, which can be found at http://www.bioinformaticsonline.org.

One aspect of this discipline that has been quite remarkable to me is the willingness of most investigators, especially the pioneers in the field, to share their results with colleagues. I have had the privilege of personally knowing several of these early investigators, especially David Lipman, Hugo Martinez (with whom I spent a sabbatical year), and Temple Smith. The tremendous accomplishments of these

people became even more meritorious because they freely shared the results of their efforts with colleagues. In doing so, they were very much responsible for the eventual success of the sequence analysis field in both the academic and commercial areas.

This large project has required much support and help. Part of this book was derived from class notes for a course in "Bioinformatics and Genome Analysis" at the University of Arizona in the 1999 and 2000 academic years. Many students made very useful suggestions and were helpful in finding errors; I want to particularly thank Bryan Zeitler for providing many corrections. Any remaining errors will be corrected on the book's Web site. I am grateful to Bill Pearson for information about the FASTA suite of programs, to Julie Thompson and John Kececioglu for comments on Chapter 4, to Steve Henikoff for reading Chapter 3, and to Michael Zuker for helpful comments on the writing of Chapter 5. Bill Montfort provided information about PDB files for Chapter 9, and Roger Miesfeld provided the example of complex gene regulation in Chapter 8. Jun Zhu was very kind in answering my questions about the Bayes block aligner for Chapter 3. My department has been most patient and supportive as I skipped meetings and seminars to complete or revise another chapter, over a period of three years. During this time, Rob Han and Juwon Kim provided the very large number of papers and book chapters that I needed on a regular basis with a very short turnaround time, allowing me more time to digest the information. My editor, Judy Cuddihy of Cold Spring Harbor Laboratory Press, guided me through the process of writing with great skill and was very patient as she tried to keep me to a reasonable writing schedule, providing needed encouragement for completing the project. Elisabeth Cuddihy checked most of the Web sites, carefully went through formulas and numerical examples, and helped to write parts of the glossary. I also thank Joan Ebert and Jan Argentine in the Development Department and Pat Barker and Denise Weiss in the Production Department at the Press.

Last, but not least, I thank my wife Jennifer Hall for her patience and understanding during the many times that book-writing took precedence over family matters.

David W. Mount

Contents

ATTGACTAGTAC

010011011100010

$P(S < x) = \exp[-e^{-x}]$

1

Historical Introduction and Overview

CONTENTS

INTRODUCTION

This chapter describes how bioinformatics has evolved into a new field of scientific investigation, describes the roles of biological and computational research in this field, and provides a brief historical account. Also provided is an overview of the chapters in this second edition. References to earlier and current reference books, articles, reviews, and journals provide a broader view of the field.

A Note about the Structure of Chapters in This Book

Each chapter will open with an introductory paragraph briefly explaining the goals of the chapter, to be followed by separate chapter guides for biologists and for computational scientists that introduce chapter concepts which may not be familiar to their respective fields and which provide orientation to the chapter topics. "What Should Be Learned in This Chapter" points out important concepts and practical topics for the chapter. A chapter glossary containing definitions for essential terms used in the chapter rounds out the introductory section of the chapter. "Search Terms from This Chapter" at the end of each chapter is a listing of terms to track current Internet addresses of sites referenced in the chapter using search engines. Each chapter concludes with a set of problems exploring the concepts and procedures described in the chapter. Because it is an introduction and overview, Chapter 1 is more general in nature than the remainder of the chapters in this book, and there is no problem set.

Book Guide for Biologists

In this chapter, we describe how DNA sequences obtained in the laboratory are usually stored in computer files much like ordinary text files. The sequence file includes information about the sequence, such as the organism, laboratory of origin, literature citations, a name of the sequence (if a known gene), and one or more unique numbers, with the main ones being sequence accession numbers or other identifiers of the particular sequence. Because there are so many sequences, the sequence files are organized into a database, e.g., GenBank, that is indexed so that specific sequences may be retrieved quite readily based on the information in each sequence file; e.g., all of the sequences for a particular organism. The database format, called a relational database, comprises cross-referenced or "keyed" tables. A computer program called a database management system is used to set up databases, place data into them, and retrieve data from them. Even though the sequences are text files, the commonly used commercial text editors should not be used to examine or manipulate these files because these programs contaminate the file by introducing other control characters into the file. Instead, there are many computer programs available for this purpose, accessible on Web sites, but also available for local, stand-alone computer systems. Described throughout this book, these programs can display or retrieve part or all of sequences from a database and store them away in a suitable format, often a new database.

The remainder of this chapter describes methods for comparing and analyzing sequences and serves as a brief survey of each of the other chapters of this book. The main concept introduced is sequence alignment, which is like trying to compare every letter and word in a sentence with those in another to determine whether any of them are in the same order, thus suggesting that they are about a similar subject. DNA and protein sequences that can be aligned in this manner have a related biological function and a common evolutionary origin. Aligning sequences is a very difficult computational problem because there are so many comparisons to be made and many different possible ways to align the sequences by including gaps. The dynamic programming algorithm is a computational method for breaking down this problem into a much smaller one in which only the first parts of the sequences are compared. Once the best alignment in an initial sequence region has been found, the comparison is progressively extended throughout the sequences. Similar approaches are also used to find the most complementary regions of all possible ones that can base-pair in RNA sequences, or to compare protein three-dimensional structures.

In the genome age, a major research goal is to find the functions of genes and to define their interactions in a particular organism. In the laboratory, genetic manipulation of model organisms, microarrays for measuring gene expression, and proteomics for analysis of proteins are used for collecting new biological data. The information obtained from these experiments needs to be organized into a suitable database format, and computer programs need to be developed

to access the information and to analyze the data, preferably through Web sites. Bioinformaticists have written many types of programs that perform manipulation and analysis of sequences.

Today, the process is simplified by writing small modular programs called "objects" that perform simple tasks. Libraries of these objects that can perform almost any task can be included in larger programs for more detailed types of sequence analyses. For example, one module may retrieve a sequence from GenBank and another may change the retrieved sequence into a particular type of sequence format. The BioPerl objects are an example of programming objects of this type. BioPerl objects can be included in programs written in the Perl programming language. Using such previously prepared objects, any application for sequence manipulation or analysis can be readily developed with only a little training or with a little help from a student or professional programmer. However, to become truly proficient, students with a biological background should strengthen their biology background through courses on the experimental methods used in molecular biology and biochemistry as well as courses in genetic analysis including population and evolutionary genetics. They should also include advanced courses in mathematics, probability and statistics, data management, data-mining and modeling tools, computer science courses in algorithm design and programming, and whatever courses are available in genome analysis and topics in bioinformatics.

Book Guide for Computational Scientists

The fundamental unit of life is the microscopic cell, which comprises a protective membrane surrounding a collection of organelles (subcellular structures) as well as large and complex molecules that provide the cell's structure and energy and enable it to reproduce. In plants and animals, individual cells cooperate to form multicellular tissues and organ systems that meet the biological needs of an organism. This book is primarily concerned with the analysis of biological sequences that regulate all biological processes in cells and organisms, including instructions for the organization of cells during the development of an organism.

The sequences are stored in very long chemical strings called DNA, which comprises four different informational letters or bases—A, G, C, and T (which stand for adenine, guanine, cytosine, and thymine)—as well as backbones of sugars and phosphates. The DNA strings are tightly wound into subcellular structures called chromosomes, which are large enough to be seen with the aid of a microscope. Except for simpler organisms like single-celled bacteria, the chromosomes are located in a visible cellular structure known as the nucleus, surrounded by cellular material known as the cytoplasm of the cell. Tissue cells have two sets of chromosomes, a type of genetic makeup referred to as diploid, with one set of chromosomes coming from one parent and the other set coming from the other. During sexual development, sexual cells called gametes (sperm or egg cells) with one set of chromosomes (haploid) are produced. The haploid collection of chromosomes in a cell comprises the genome of an organism. During sexual reproduction, the sequences (genes) on the two parental chromosomes are reassorted. Gametes with one set of the newly assorted chromosomes are then produced and eventually passed on to the next generation.

Regions of DNA sequence along a chromosome encode instructions for the manufacture of proteins in the cell. Proteins are linear chains of a set of 20 chemically active building blocks known as amino acids. Each protein has a unique sequence of amino acids that is determined by a DNA sequence on the chromosomes. The complement of proteins enables an organism to build the structures and to carry out the biological functions of that organism. Using specific biological machinery in a process called transcription, the cell "reads" the DNA sequences by searching for specific sequence patterns (such as promoters) that mark the beginning of the unit of hereditary information—the gene. Starting at that point and going in a particular chemical direction along the molecule, the sequence is read until another pattern indicating the end of the gene is reached. Transcription produces another long chemical string, called messenger RNA or mRNA, whose sequence specifies the amino acid sequence of the protein to be produced. mRNA molecules are very similar structurally and chemically to DNA except that they are usually single-stranded, they have a new base uracil (U) instead of thymine (T), and there is a different sugar in the backbone of the chain. mRNA molecules also have a specified sequence pattern that indicates where the code for the protein begins. Large organelles in the cell's cytoplasm called ribosomes bind to these start sites and, moving in a defined chemical direction, read three base positions (a codon) at a time that specify an amino acid. The corresponding amino acid for that codon is then added to a growing chain of amino acids that

comprise the protein. Amino acids are added until one of several stop codons is reached. Each protein sequence is thus generally colinear with the original codon sequence on the chromosome.

Once formed, proteins rapidly fold from a linear string into simple helical and stranded elements (secondary structures) and then organize these elements into a unique three-dimensional structure. The resulting protein molecule may serve as a tissue building block or it may have a very specific chemical activity. The collection of proteins produced by an organism, called the proteome, is responsible for the structure and biological behavior of that organism. Not all genes are translated into proteins—some are kept as four-letter sequences in RNA molecules that regulate many important cellular processes. Simple organisms have several thousand genes, and more complex ones have 12,000–35,000 genes. Some organisms, especially some plants, may have large numbers of duplicated genes or duplicated chromosomes, and these extra gene sets are copied from one generation to the next, sometimes producing genomes of 100,000 or more genes.

Fundamentally important to understanding how biologists think about genes is the concept that all organisms appear to be related through an evolutionary process. Even very different organisms, such as single-celled bacteria and multicellular plants and animals, share some of the same genes. These genes are usually not made over and over again from a new biological starting point, but have instead been copied from an ancestor gene; the DNA sequence is then changed randomly to a limited degree by mutation. Organisms can be grouped into binary trees in which outer branches represent more closely related, more recently evolved, organisms, and inner branches represent the more primitive, usually simpler, organisms. These trees can be based on biological information (structure or behavior of organisms), but more recently, with the sequencing of the genomes of model organisms, they are based on the complement of genes found in genomes. As a result, much of the activity in bioinformatics is concerned with comparing genes and the encoded proteins as strings of letters from an alphabet of 4 (DNA) or 20 (protein) characters. If a protein encoded by one organism can be readily aligned with a protein encoded by another, this result tells the biologist that these genes came from a common source and have the same biological function.

There is also a second important concept to learn about genes and proteins. Instead of developing an entirely new gene to develop a new biological function, organisms appear to use three tactics. First, copies of existing genes are made and, through a process of random mutation, one of these copies can gradually change to develop a new biological activity. In those rare cases in which this new activity is of benefit to the organism, natural selection acts to establish the gene in that organism. This process can develop a whole family of related but functionally important genes that can be passed down the evolutionary tree. The second tactic is for organisms to produce new genes by combining parts (domains) of existing genes that represent an elementary unit of three-dimensional protein structure or biological function. Bioinformatics has been very much involved in the discovery of these gene and protein families and sequence domains through sequence alignment methods and searching for common patterns in sequences using statistical methods. The third method of diversifying the function of existing genes in a given organism is by the rare transfer of genes from a different, unrelated organism. Normally, genes are transferred from parent to offspring, a process that biologists call vertical transfer because the inheritance pattern follows the branches of a tree-like pedigree. In another type of inheritance, a foreign piece of DNA is accidentally transferred into a cell and incorporated into the existing chromosome, thus adding more genes. This type of inheritance is called horizontal transfer because the contributing organisms are from different species that do not normally exchange DNA. In ancient life, even whole cells were thought to merge to create new organisms. Horizontal transfer continues to play an important role in the evolution of the single-celled bacteria, particularly in the development of resistance to specific antibiotics. Thus, all organisms share many genes or parts of genes, and these can now be identified in the genomes of these organisms.

Two challenges facing biologists are to discover the biological function of genes and to understand how the interaction of genes regulates biological processes. If there is no information available from studies of the homologous gene in another organism, then several experimental approaches can be used: (1) The sequence of the gene may be disrupted (a sequence change or an insertion or deletion of sequence) or (2) the expression of the gene may be turned off by introducing an altered form of the mRNA copy of the gene, which can lead to degradation of the cell's mRNA copies and also change the structure of the gene in a semi-permanent, heritable way (epigenetic change) so that the gene is no longer transcribed by the cell. Gene function is inferred from the effect of these genetic or epigenetic changes using a battery of biological tests. In addition to genetic analysis, biologists also use methods to follow the expression of most cell genes and proteins using DNA microarray technologies and protein analysis experiments (proteomics). These methods support the discovery of patterns of gene expression and protein occurrence that relate to biological function such as, for example, abnormal patterns in cancer cells.

BASICS FOR TRAINING STUDENTS IN BIOINFORMATICS

- Know where sequence, protein, and genome data are located and how they are stored.
- Have the programming skills needed to retrieve data from external Web sites and databases and then to organize and store the data in a suitable database format.
- Have sufficient biological knowledge to relate biological information on gene function and protein structure/function with sequence and genome information and know how to store these data in a suitable format.
- Stay current on methods of sequence and genome analysis for finding related genes of similar function in different organisms, for analysis of gene regulation, and for prediction of protein structure and function.
- Know how to manage and use large data sets such as gene expression microarrays, proteomics data, and sequence variations in populations.
- Be able to integrate diverse data sets and use existing tools for mining these data to discover new relationships among the data.
- Be conversant with computer scientists, mathematicians, and statisticians to develop additional analytical tools and models as needed.

Glossary Terms

Alignment is the procedure of comparing two or more sequences by looking for a series of individual characters or character patterns that are in the same order in the sequences.

Algorithm is a method for data analysis that can be proven mathematically to produce a desired solution, with emphasis on a time- and space-efficient design. An example is the dynamic programming algorithm used for sequence alignment.

Annotation is locating genes in a genome sequence, including protein-encoding and RNA-encoding genes, thus providing the sequence and location of the encoded proteins and RNA molecules.

Codon is a length of three nucleotides in DNA that is translated by the cell as an amino acid position in a protein. Of the 64 possible codons, 61 are usually read as one of the 20 amino acids, and the remaining 3 are read as stop codons indicating the end of the protein. mRNAs carry the information for protein sequences as a sequence of codons.

Comparative genomics is a comparison of gene numbers, gene locations, and biological functions of genes in the genomes of diverse organisms, one objective being to identify families of genes that play a unique biological role in a particular organism.

Database is an organized system for storing data, usually in a set of cross-referenced (cross-keyed) tables called a relational database.

Distance score refers to the positions in a sequence alignment at which the sequence characters are different or rarely found in alignments of related sequences and usually, but not always, does not include gapped positions.

DNA is a double-stranded, helical molecule comprising a sequence of four nucleotides—A (adenine), G (guanine), C (cytosine), and T (thymine)—in each strand. A in one strand is always paired with T in the other, and G is always paired with C. The chemical interaction of these nucleotide pairs holds the strands together. Using these A/T and G/C pairing rules, a new strand is synthesized by separating the strands and using each single strand as a template for a new one. This molecular mechanism, discovered by James D. Watson and Francis Crick, underlies all biological reproduction. The sequence of nucleotides in DNA molecules can be read by DNA sequencing machines 500–800 at a time.

Dot matrix analysis provides a graphical method for comparing two sequences. One sequence is written horizontally across the top (or bottom) of the graph and the other along the left-hand side. Dots are placed within the graph at the intersection of the same letter appearing in both sequences. Diagonal rows indicate similarity.

Dynamic programming is an algorithm used for aligning sequences allowing for matches, mismatches, and gaps (insertions or deletions). The algorithm first finds the best alignment at the beginning of the sequences and then sequentially adds more of each sequence until both sequences have been aligned.

Extreme value distribution Some measurements such as sequence alignment scores are found to follow a distribution that has a long tail which decays at high values much more slowly than found in a normal distribution. One slow-falling type is called the extreme value distribution. The alignment scores between unrelated or random sequences are an example. These scores can reach very high values, particularly when a large number of comparisons are made, as in a database similarity search. The probability of a particular score may be accurately predicted by the extreme value distribution, which follows a double-negative exponential function after Gumbel.

Functional genomics is an assessment of the function of genes identified by between-genome comparisons. The function of a newly identified gene is commonly tested by introducing mutations into the gene and then examining the resulting organism for altered properties.

Gene is a length of DNA that specifies a unit of biological function, usually the amino acid sequence of a protein. The DNA is copied into mRNA molecules using pairing rules like those used to synthesize new DNA strands.

Gene expression microarrays are arrays on a microscope slide of tiny spots of DNA representing a large proportion of all the genes of an organism. Microarrays are used for comparing the levels of mRNA of the entire set of genes in a biological sample. Complementary DNAs (cDNAs) of the chemically unstable mRNAs are synthesized and labeled with a fluorescent dye to detect the mRNAs.

Genome is the entire complement of genetic material of an organism including all of the genes that specify proteins and RNA molecules, and any other sequences that are present. This comprises half the paired chromosomes in a somatic (body) cell and all of the chromosomes in a germ (sex) cell.

Global sequence alignment is an alignment method that deliberately includes all of the sequences in the alignment.

Homologous describes genes that have arisen from a common ancestor gene, as evidenced by their having similar sequences.

Local sequence alignment is an alignment method that aligns regions of sequences by the highest density of matches.

Maximum parsimony tree is a graphical representation of the observed changes in a multiple sequence alignment and is done in such a way that the sum of the number of changes along the branches of the tree is a minimum.

Model organism is an organism that can be manipulated genetically to discover the function of a particular gene by observing any biological changes in the resulting organism. The results obtained from such model organism studies can often be applied to other organisms. Examples of model organisms are the fruit fly *Drosophila*, yeast, zebrafish, mouse, and the plant *Arabidopsis*.

Motif (protein) is usually a short pattern of amino acids that can represent an active site or functional region in a protein structure.

Multiple sequence alignment is an alignment of three or more sequences that attempts to place sequence positions related by function and evolution in the same column of the alignment allowing for mismatches and gaps (deletions or insertions).

Orthologous describes genes found in two or more different organisms which are so uniquely and strikingly similar that they are predicted to have the same biological function in those organisms.

Orthologs are genes found to be orthologous.

Paralogous describes a family of similar genes in an organism that are predicted to have arisen by gene duplication in ancestors of the organism.

Paralogs are genes that are found to be paralogous.

Phylogenetic analysis of sequences attempts to discover evolutionary relationships among a set of similar sequences by organizing them into a tree representation with placement of more similar sequences on neighboring branches.

Position-specific scoring matrix (PSSM) is a table that represents the variation found in the columns of a multiple sequence alignment of a set of related sequences. The table columns correspond to the columns in the alignment and the rows represent the frequency of the sequence characters in each column. The frequencies are often divided by the frequencies of the characters in the sequences in order to create odds scores. For convenience, odds scores are often turned into log odds scores.

Protein is a molecule comprising a long chain of amino acids, the sequence of which is specified by the sequence of codons in a gene. The chain folds into a three-dimensional structure unique to a particular protein that has biological activity.

Proteome is the entire complement of proteins synthesized by an organism.

Proteomics is the analysis of biological samples for protein content, modification, and activity.

RNA is usually a single-stranded molecule and, like DNA, comprises four nucleotides—A, G, C, and U (uracil). RNA is produced by copying one of the two strands of a DNA molecule in the 5′ to 3′ chemical direction. Messenger RNAs (mRNAs) copy the information from DNA for protein sequences, and this information is subsequently translated by the cell into protein sequences.

Sequence similarity Two sequences are similar if the order of sequence characters is recognizably the same in the sequences, and is usually found by showing that they can be aligned.

Similarity score (sequence alignment) is the sum of the number of identical matches and conservative (high scoring) substitutions in a sequence alignment divided by the total number of aligned sequence characters. Gaps are usually, but not always, ignored.

Synteny refers to the conserved order of genes in related organisms due to their derivation from a common ancestor.

WHAT IS BIOINFORMATICS?

Previously, bioinformatics was defined as an interdisciplinary field involving biology, computer science, mathematics, and statistics to analyze biological sequence data, genome content, and arrangement, and to predict the function and structure of macromolecules. With the advent of the genome era, bioinformatics now plays added roles in biological and medical research and accounts for an increasing number of publications each year (see Luscome et al. 2001).

Bioinformatics overlaps other areas of research that are designated informatics and computational biology. Informatics has traditionally been a discipline in which mathematicians, computer scientists, statisticians, and engineers develop technologies for supporting information management in fields like health care. Bioinformatics is now involved in these activities by organizing biological data related to genomes with a view to applying this information in agriculture, pharmacology, and other commercial applications. The biological information is of two types—sequence information and content obtained from genomes and structure–function analysis of the gene products obtained by experiment. In the case of the human genome, this role of bioinformatics means collecting information about the biological function of the ~35,000 human genes to discover which ones have the most significant role in disease. Using modern technologies, such as gene expression microarrays, genetic manipulation of genes in cells and organisms, and rapid structural and functional assessment of proteins, many new data on gene function are being assembled.

The closely related field of computational biology also provides computational support for many of the above purposes, but there are differences with the support provided by bioinformatics. On the one hand, computational biology generally is concerned with the development of novel and efficient algorithms that can be proven to work on a difficult problem, such as multiple sequence alignment or genome fragment assembly. On the other hand, bioinformatics focuses more on the development of practical tools for data management and analysis, e.g., display of genome information and sequence analysis, but with less emphasis on efficiency and proven accuracy. In many cases, such as multiple sequence alignment and database similarity searching, a suitable model for sequence change is not available, or a reasonably defined sequence analysis problem is too complex to solve in a reasonable period of time. In such cases, bioinformatics can provide computational methods that meet current needs but lack a provable foundation. Thus, today the field of bioinformatics supports a broad spectrum of research that includes determining the biological significance of the data, provides the expertise to organize it, and develops practical computational tools needed to mine the data for new information. When the information is of practical importance, bioinformatics assists with practical applications such as the identification of new protein targets for drug therapy. Admittedly, however, the fields of bioinformatics, informatics, and computational biology are rapidly evolving and will play changing roles in the use of genome data.

THE FIRST SEQUENCES TO BE COLLECTED WERE THOSE OF PROTEINS

The development of protein sequencing methods (Sanger and Tuppy 1951) led to the sequencing of representatives of several of the more common protein families, such as cytochromes from a variety of organisms. Margaret Dayhoff (1972, 1978) and her collaborators at the National Biomedical Research Foundation (NBRF), Washington, D.C., were the first to assemble databases of these sequences into a protein sequence atlas in the 1960s, and their collection center eventually became known as the Protein Information Resource (PIR, formerly Protein Identification Resource; http://watson.gmu.edu:8080/pirwww/index.html). The NBRF maintained the database from 1984, and in 1988, the PIR-International Protein Sequence Database (http://www-nbrf.georgetown.edu/pir) was established as a collaboration of NBRF, the Munich Center for Protein Sequences (MIPS), and the Japan International Protein Information Database (JIPID).

Margaret Dayhoff

Dayhoff and her coworkers organized the proteins into families and superfamilies based on the degree of sequence similarity. Tables that reflected the frequency of changes observed in the sequences of a group of closely related proteins were then derived. As protein sequences become more varied, one has to be concerned as to whether a given amino acid may have changed more than one time. To avoid this problem of multiple changes, proteins that were less than 15% different were chosen to avoid the chance that the observed amino acid changes reflected two sequential amino acid changes instead of only one. From aligned sequences, a phylogenetic tree was derived showing graphically which sequences were most related and therefore shared a common branch on the tree. Once these trees were made, they were used to score the amino acid changes that occurred during evolution of the genes for these proteins in the various organisms from which they originated (Fig. 1.1).

The rule used in evaluating the results of phylogenetic tree analysis is that the more identical and conserved amino acids that there are in two sequences, the more likely they are to have been derived from a common ancestor gene during evolution. If the sequences are very much alike, the proteins probably have the same biochemical function and three-dimensional structural fold. Thus, a set of matrices (tables), called PAM (percent accepted mutation) tables for the percent amino acid mutations accepted by evolutionary selection, were constructed. These tables showed the probability that one amino acid had changed into any other in these trees, thus indicating which amino acids are most conserved at the corresponding position in two sequences. PAM tables are still used to measure similarity between protein sequences and are used in database searches to find sequences that match a query sequence.

Dayhoff and her colleagues contributed in several ways to modern biological sequence analysis by providing the first protein sequence database as well as PAM tables for performing protein sequence comparisons. Amino acid substitution tables are routinely used in performing sequence alignments and database similarity searches, and their use for this purpose is discussed in Chapters 3 and 6.

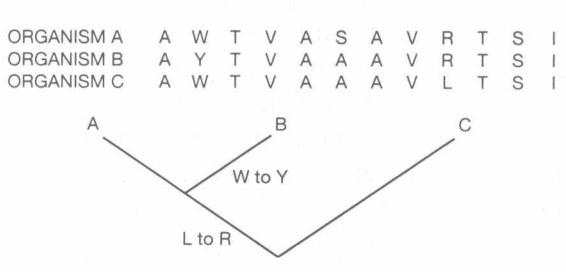
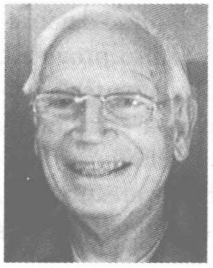


FIGURE 1.1. Method of predicting phylogenetic relationships and probable amino acid changes during the evolution of related protein sequences. Shown are three highly conserved sequences (A, B, and C) of the same protein from three different organisms. The sequences are so similar that each position should only have changed once during evolution. The proteins differ by one or two substitutions, allowing the construction of the tree shown. Once this tree is obtained, the indicated amino acid changes can be determined. The particular changes shown are examples of two that occur much more often than expected by a random replacement process.

DNA SEQUENCE DATABASES DATE FROM THE EARLY 1980s

Collecting DNA sequences into the GenBank database was initiated by the Theoretical Biology and Biophysics Group founded in 1974 by George I. Bell at Los Alamos National Laboratory in New Mexico. This group of physicists wanted to provide a theoretical background to experimental work, primarily in immunology. The

group's research expanded into other areas of computational biology and bioinformatics and into the development of the first versions of the GenBank under the leadership of Walter Goad and colleagues between 1982 and 1992. Goad first conceived of the GenBank prototype in 1979. Translated DNA sequences were also included in the Protein Information Resource (PIR) database at the National Biomedical Research Foundation in Washington, D.C. The European Molecular Biology Laboratory (EMBL) Data Library was founded in 1980 (http://www.ebi.ac.uk), and in 1984 the DNA DataBank of Japan (DDBJ), Mishima, Japan, came into existence (http://www.ddbj.nig.ac.jp). GenBank is now under the auspices of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). GenBank, EMBL, and DDBJ have now formed the International Nucleotide Sequence Database Collaboration (http://www.ncbi.nlm.nih.gov/collab), which acts to facilitate exchange of data on a daily basis. PIR has made similar arrangements.



Walter Goad

Many types of sequence databases are described in the first issue each year of the journal *Nucleic Acids Research*. The growth of the number of sequences in GenBank can be tracked at http://www.ncbi.nlm.nih.gov/GenBank/genebankstats.html.

Initially, a sequence entry in these databases included a computer file name and DNA or protein sequence files. The entries were eventually expanded to include much more information about the sequence, such as function, mutations, encoded proteins, regulatory sites, and references. This annotated information was then placed along with the sequence into a database format that could be readily searched for many types of information. There are many such databases and formats, which are discussed in Chapter 2.

The number of entries in the nucleic acid sequence databases GenBank and EMBL has continued to increase enormously from the daily updates. Annotating all of these new sequences is a time-consuming, painstaking, and sometimes error-prone process. As time passes, the process is becoming more automated, creating additional problems of accuracy and reliability. GenBank grew from 1.26×10^9 bases in December, 1997 to 39×10^9 bases in April, 2004. Despite the exponentially increasing numbers of sequences stored, the implementation of efficient search methods has provided ready public access to these sequences.

To decrease the number of matches to a database search, nonredundant databases that list only a single representative of identical sequences have been prepared. An example of this is the NCBI RefSeq database. However, many sequence databases, such as the nonredundant NR databases used for BLAST search at NCBI, still include a large number of entries of the same gene or protein sequences originating from sequence fragments, patents, replica entries from different databases, and other such sequences.

SEQUENCES CAN BE EASILY RETRIEVED FROM PUBLIC DATABASES

An important step in providing sequence database access was the development of Web pages that allow queries to be made of the major sequence databases (GenBank, EMBL, etc.). An early example of this technology at NCBI was a menu-driven program called GENINFO developed by D. Benson, D. Lipman, and colleagues. This program searched rapidly through previously indexed sequence databases for entries that matched a biologist's query. Subsequently, a derivative program called Entrez (http://www.ncbi.nlm.nih.gov/Entrez) with a simple window-based interface, and eventually a Web-based interface, was developed at NCBI.

David Lipman

The idea behind these programs was to provide an easy-to-use interface with a flexible search procedure to the sequence databases using keywords searching on standardized entry fields. Sequence entries in the major databases have additional information about the sequence included with the sequence entry, such as accession or index number, name and alternative names for the sequence, names of relevant genes, types of regulatory sequences, the source organism, references, and known mutations. Entrez accesses this information, thus allowing rapid searches of entire sequence databases for matches to one or more specified search terms.

These programs also can locate similar sequences