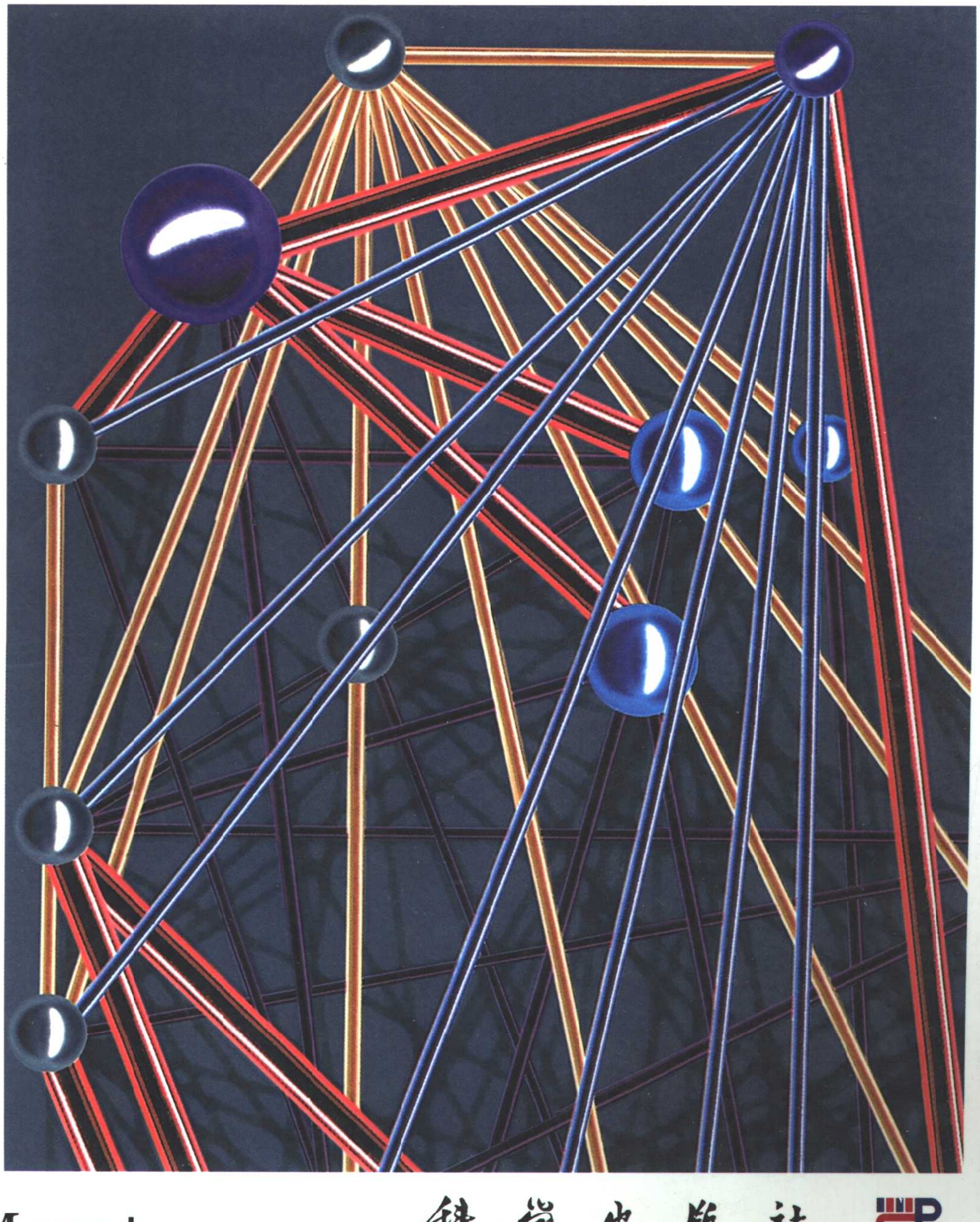
Bioinformatics

生物信息学

(影印版)

序列与基因组分析

Sequence and Genome Analysis



David W. Mount

科学出版社

Cold Spring Harbor Laboratory Press

生物信息学

序列与基因组分析

Bioinformatics

Sequence and Genome Analysis

David W. Mount

科学出版社

2002

内容简介

生物信息学是将计算和分析方法应用于研究DNA、蛋白质等生物学问题的一个迅速发展的领域，是本世纪生物科学的主要前沿之一。本书继承了冷泉港实验室出版物一贯的权威、经典的风格，全面系统地介绍了这方面的理论与应用。全书深入浅出、图文并茂、资料丰富、内容新颖，是从事和即将从事生物信息学研究的科研人员、技术人员、研究生的重要参考书和入门书。

David W. Mount

Bioinformatics: Sequence and Genome Analysis

图字：01-2002-1681

图书在版编目（CIP）数据

生物信息学：序列与基因组分析/（美）蒙特（Mount，D. W.）著．—影印版．—北京：科学出版社，2002.6
ISBN 7-03-010441-2

Ⅰ．生…　Ⅱ．蒙…　Ⅲ．生物信息论-英文　Ⅳ．Q811.4

中国版本图书馆CIP数据核字（2002）第030915号

科学出版社　出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
新蕾印刷厂　印刷
科学出版社发行　各地新华书店经销
*
2002年6月第　一　版　　开本：890×1240　1/16
2002年6月第一次印刷　　印张：36
印数：1—3 000　　字数：1 129 000

定价：73.00元

（如有印装质量问题，我社负责调换〈路通〉）

*This book is dedicated to the following individuals who have
contributed much to the field of sequence analysis:
David Lipman,
Bill Pearson,
Temple Smith,
and Michael Waterman
and to the memory of
Margaret Dayhoff and Walter Goad*

The *Bioinformatics* Web Site

Access to the On-line Text and Associated Resources

This print edition of *Bioinformatics* is associated with a Web site (www.bioinformaticsonline.org) that will add to and extend the contents of the book.

When the site is launched, registered purchasers of the book will be able to (at no extra charge):

- Access Web sites referred to in the text.
- Access problem sets for classroom use and other useful material not included in the print edition.
- Receive E-mail alerts about peer-reviewed, new, and updated information that extends the scope and content of the book.

To register at www.bioinformaticsonline.org:

1. Open the home page of the site.
2. Follow the registration procedure that begins on that page.
3. When prompted, enter the unique access code that is printed on the inside front cover of this book.
4. When prompted, enter your E-mail address as your user name and a password of your choice.
5. Complete the registration procedure as requested.

The Web site contains answers to FAQs about the registration procedure and a demonstration of the functions available to registered users. For additional assistance with registration, and for all other inquiries about the *Bioinformatics* Web site, please use the E-mail addresses provided at the site.

Preface

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 over-simplifications 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 have not 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

CHAPTER 1

Historical Introduction and Overview

THE DEVELOPMENT OF SEQUENCE ANALYSIS METHODS has depended on the contributions of many individuals from varied scientific backgrounds. This chapter provides a brief historical account of the more significant advances that have taken place, as well as an overview of the chapters of this book. Because many contributors cannot be mentioned due to space constraints, additional references to earlier and current reference books, articles, reviews, and journals provide a broader view of the field and are included in the reference lists to this chapter.

THE FIRST SEQUENCES TO BE COLLECTED WERE THOSE OF PROTEINS

Margaret Dayhoff

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, DC, 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).

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. 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).

ORGANISM A	A	W	T	V	A	S	A	V	R	T	S	I
ORGANISM B	A	Y	T	V	A	A	A	V	R	T	S	I
ORGANISM C	A	W	T	V	A	A	A	V	L	T	S	I

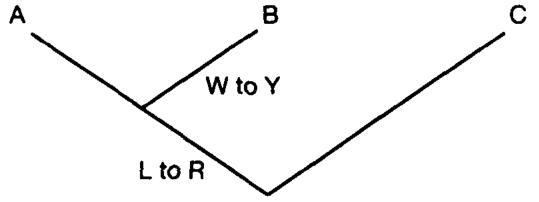
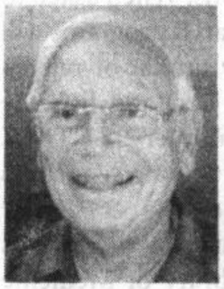


Figure 1.1. Method of predicting phylogenetic relationships and probable amino acid changes during the evolution of related protein sequences. Shown are three highly conserved sequences (A, B, and C) of the same protein from three different organisms. The sequences are so similar that each position should 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.

Subsequently, a set of matrices (tables)—the percent amino acid mutations accepted by evolutionary selection or PAM tables—which showed the probability that one amino acid changed into any other in these trees was constructed, thus showing which amino acids are most conserved at the corresponding position in two sequences. These tables are still used to measure similarity between protein sequences and in database searches to find sequences that match a query sequence. The rule used 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 folds. Thus, 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 7.

DNA SEQUENCE DATABASES



Walter Goad

DNA sequence databases were first assembled at Los Alamos National Laboratory (LANL), New Mexico, by Walter Goad and colleagues in the GenBank database and at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. Translated DNA sequences were also included in the Protein Information Resource (PIR) database at the National Biomedical Research Foundation in Washington, DC. Goad had conceived of the GenBank prototype in 1979; LANL collected GenBank data from 1982 to 1992. GenBank is now under the auspices of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). The EMBL Data Library was founded in 1980 (http://www.ebi.ac.uk). In 1984 the DNA DataBank of Japan (DDBJ), Mishima, Japan, came into existence (http://www.ddbj.nig.ac.jp). 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.

Many types of sequence databases are described in the first annual issue 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 included a computer filename and DNA or protein sequence files. These were eventually expanded to include much more information about the sequence, such as function, mutations, encoded proteins, regulatory sites, and references. This 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. In December 1997, there were 1.26×10^9 bases in GenBank; this number increased to 2.57×10^9 bases as of April 1999, and 1.0×10^{10} as of September 2000. 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, non-redundant databases that list only a single representative of identical sequences have been prepared. However, many sequence databases 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.

SEQUENCE RETRIEVAL FROM PUBLIC DATABASES

David Lipman

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. The idea behind these programs was to provide an easy-to-use interface with a flexible search procedure to the sequence databases.

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 (called "neighbors" by ENTREZ) on the basis of previous similarity comparisons. When asked to perform a search for one or more terms in a database, simple pattern search programs will only find exact matches to a query. In contrast, ENTREZ searches for similar or related terms, or complex searches composed of several choices, with great ease and lists the found items in the order of likelihood that they matched the original query. ENTREZ originally allowed straightforward access to databases of both DNA and protein sequences and their supporting references, and even to an index of related entries or similar sequences in separate or the same databases. More recently, ENTREZ has provided access to all of Medline, the full bibliographic database of the National Library of Medicine (NLM), Washington, DC. Access to a number of other databases, such as a phylogenetic database of organisms and a protein structure database, is also provided. This access is provided without cost to any user—private, government, industry, or research—a decision by the staff of NCBI that has provided a stimulus to biomedical research that cannot be underestimated. NCBI presently handles several million independent accesses to their system each day.

A note of caution is in order. Database query programs such as ENTREZ greatly facilitate keeping up with the increasing number of sequences and biomedical journals. However, as with any automated method, one should be wary that a requested database search may not retrieve all of the relevant material, and important entries may be missed. Bear in mind that each database entry has required manual editing at some stage, giving rise to a low frequency of inescapable spelling errors and other problems. On occasion, a particular reference that should be in the database is not found because the search terms may be misspelled in the relevant database entry, the entry may not be present in the database, or there may be some more complicated problem. If exhaustive and careful attempts fail, reporting such problems to the program manager or system administrator should correct the problem.

SEQUENCE ANALYSIS PROGRAMS

Because DNA sequencing involves ordering a set of peaks (A, G, C, or T) on a sequencing gel, the process can be quite error-prone, depending on the quality of the data.

Methods for DNA sequencing were developed in 1977 by Maxam and Gilbert (1977) and Sanger et al. (1977). They are described in greater detail at the beginning of Chapter 2.

As more DNA sequences became available in the late 1970s, interest also increased in developing computer programs to analyze these sequences in various ways. In 1982 and 1984, *Nucleic Acids Research* published two special issues devoted to the application of computers for sequence analysis, including programs for large mainframe computers down to the then-new microcomputers. Shortly after, the Genetics Computer Group (GCG) was started at the University of Wisconsin by J. Devereux, offering a set of programs for analysis that ran on a VAX computer. Eventually GCG became commercial (http://www.gcg.com/). Other companies offering microcomputer programs for sequence analysis, including Intelligenetics, DNAStar, and others, also appeared at approximately the same time. Laboratories also developed and shared computer programs on a no-cost or low-cost basis. For example, to facilitate the collection of data, the programs PHRED (Ewing and Green 1998; Ewing et al. 1998) and PHRAP were developed by Phil Green and colleagues at the University of Washington to assist with reading and processing sequencing data. PHRED and PHRAP are now distributed by CodonCode Corporation (http://www.codoncode.com).

These commercial and noncommercial programs are still widely used. In addition, Web sites are available to perform many types of sequence analyses; they are free to academic institutions or are available at moderate cost to commercial users. Following is a brief review of the development of methods for sequence analysis.

THE DOT MATRIX OR DIAGRAM METHOD FOR COMPARING SEQUENCES

In 1970, A.J. Gibbs and G.A. McIntyre (1970) described a new method for comparing two amino acid and nucleotide sequences in which a graph was drawn with one sequence written across the page and the other down the left-hand side. Whenever the same letter appeared in both sequences, a dot was placed at the intersection of the corresponding sequence positions on the graph (Fig. 1.2). The resulting graph was then scanned for a series of dots that formed a diagonal, which revealed similarity, or a string of the same characters, between the sequences. Long sequences can also be compared in this manner on a single page by using smaller dots.

The dot matrix method quite readily reveals the presence of insertions or deletions between sequences because they shift the diagonal horizontally or vertically by the amount of change. Comparing a single sequence to itself can reveal the presence of a repeat of the same sequence in the same (direct repeat) or reverse (inverted repeat or palindrome) orientation. This method of self-comparison can reveal several features, such as similarity between chromosomes, tandem genes, repeated domains in a protein sequence, regions of low sequence complexity where the same characters are often repeated, or self-complementary sequences in RNA that can potentially base-pair to give a double-stranded structure. Because diagonals may not always be apparent on the graph due to weak similarity, Gibbs and McIntyre counted all possible diagonals and these counts were compared to those of random sequences to identify the most significant alignments.

Maizel and Lenk (1981) later developed various filtering and color display schemes that greatly increased the usefulness of the dot matrix method. This dot matrix representation of sequence comparisons continues to play an important role in analysis of DNA and protein sequence similarity, as well as repeats in genes and very long chromosomal sequences, as described in Chapter 3 (p. 59).

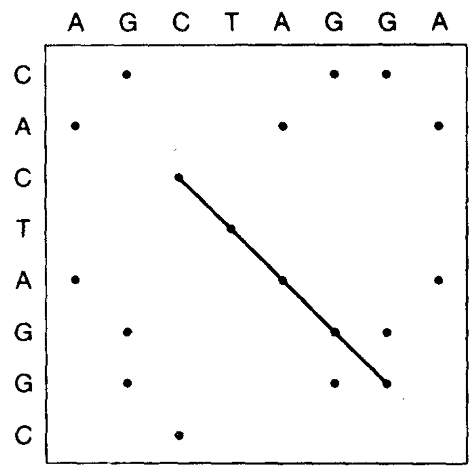


Figure 1.2. A simple dot matrix comparison of two DNA sequences, AGCTAGGA and GACTAGGC. The diagonal of dots reveals a run of similar sequence CTAGG in the two sequences.

ALIGNMENT OF SEQUENCES BY DYNAMIC PROGRAMMING

Although the dot matrix method can be used to detect sequence similarity, it does not readily resolve similarity that is interrupted by regions that do not match very well or that are present in only one of the sequences (e.g., insertions or deletions). Therefore, one would like to devise a method that can find what might be a tortuous path through a dot matrix, providing the very best possible alignment, called an optimal alignment, between the two sequences. Such an alignment can be represented by writing the sequences on successive lines across the page, with matching characters placed in the same column and unmatched characters placed in the same column as a mismatch or next to a gap as an insertion (or deletion in the other sequence), as shown in Figure 1.3. To find an optimal alignment in which all possible matches, insertions, and deletions have been considered to find the best one is computationally so difficult that for proteins of length 300, 10^{88} comparisons will have to be made (Waterman 1989).

To simplify the task, Needleman and Wunsch (1970) broke the problem down into a progressive building of an alignment by comparing two amino acids at a time. They started at the end of each sequence and then moved ahead one amino acid pair at a time, allowing for various combinations of matched pairs, mismatched pairs, or extra amino acids in one sequence (insertion or deletion). In computer science, this approach is called dynamic programming. The Needleman and Wunsch approach generated (1) every possible alignment, each one including every possible combination of match, mismatch, and single insertion or deletion, and (2) a scoring system to score the alignment. The object was to determine which was the best alignment of all by determining the highest score. Thus, every match in a trial alignment was given a score of 1, every mismatch a score of 0, and individual gaps a penalty score. These numbers were then added across the alignment to

SEQUENCE A	A	G	Δ	Δ	C	D	E	V	I	G
SEQUENCE B	A	G	E	Y	C	D	Δ	I	I	G

Figure 1.3. An alignment of two sequences showing matches, mismatches, and gaps (Δ). The best or optimal alignment requires that all three types of changes be allowed.

obtain a total score for the alignment. The alignment with the highest possible score was defined as the optimal alignment.

The procedure for generating all of the possible alignments is to move sequentially through all of the matched positions within a matrix, much like the dot matrix graph (see above), starting at those positions that correspond to the end of one of the sequences, as shown in Figure 1.4. At each position in the matrix, the highest possible score that can be achieved up to that point is placed in that position, allowing for all possible starting points in either sequence and any combination of matches, mismatches, insertions, and deletions. The best alignment is found by finding the highest-scoring position in the graph, and then tracing back through the graph through the path that generated the highest-scoring positions. The sequences are then aligned so that the sequence characters corresponding to this path are matched.

	G	A	T	C	T	A	
G	1						
A		2				1	
T			3		1		
C				4			
A		1				5	(minus gap penalty)

Deduced alignment with gap Δ

G A T C T A
G A T C Δ A

Figure 1.4. Simplified example of Needleman-Wunsch alignment of sequences GATCTA and GATCA. First, all matches in the two sequences are given a score of 1, and mismatches a score of 0 (not shown), chosen arbitrarily for this example. Second, the diagonal 1s are added sequentially, in this case to a total score of 4. At this point the row cannot be extended by another match of 1 to a total score of 5. However, an extension is possible if a gap is placed in GATCA to produce GATC Δ A, where Δ is the gap. To add the gap, a penalty score is subtracted from the total match score of 5 now appearing in the last row and column. The best alignment is found starting with the sequence characters that correspond to the highest number and tracing back through the positions that contributed to this highest score.

FINDING LOCAL ALIGNMENTS BETWEEN SEQUENCES

Mike Waterman

Temple Smith

The above method finds the optimal alignment between two sequences, including the entirety of each of the sequences. Such an alignment is called a global alignment. Smith and Waterman (1981a,b) recognized that the most biologically significant regions in DNA and protein sequences were subregions that align well and that the remaining regions made up of less-related sequences were less significant. Therefore, they developed an important modification of the Needleman-Wunsch algorithm, called the local alignment or Smith-Waterman (or the Waterman-Smith) algorithm, to locate such regions. They also recognized that insertions or deletions of any size are likely to be found as evolutionary changes in sequences, and therefore adjusted their method to accommodate such changes. Finally, they provided mathematical proof that the dynamic programming method is guaranteed to provide an optimal alignment between sequences. The algorithm is discussed in detail in Chapter 3 (p. 64).

Two complementary measurements had been devised for scoring an alignment of two sequences, a similarity score and a distance score. As shown in Figure 1.3, there are three types of aligned pairs of characters in each column of an alignment—identical matches, mismatches, and a gap opposite an unmatched character. Using as an example a simple scoring system of 1 for each type of match, the similarity score adds up all of the matches in the aligned sequences, and divides by the sum of the number of matches and mismatches (gaps are usually ignored). This method of scoring sequence similarity is the one most familiar to biologists and was devised by Needleman and Wunsch and used by Smith and Waterman. The other scoring method is a distance score that adds up the number of substitutions required to change one sequence into the other. This score is most useful for making predictions of evolutionary distances between genes or proteins to be used for phylogenetic (evolutionary) predictions, and the method was the work of mathematicians, notably P. Sellers. The distance score is usually calculated by summing the number of mismatches in an alignment divided by the total number of matches and mismatches. The calculation represents the number of changes required to change one sequence into the other, ignoring gaps. Thus, in the example shown in Figure 1.3, there are 6 matches and 1 mismatch in an alignment. The similarity score for the alignment is $6/7 = 0.86$ and the distance score is $1/7 = 0.14$, if the required condition is given a simple score of 1. With this simple scoring scheme, the similarity and distance scores add up to 1. Note also the equivalence that the sum of the sequence lengths is equal to twice the number of matches plus mismatches plus the number of deletions or insertions. Thus, in our example, the calculation is $8 + 9 = 2 \times (6 + 1) + 3 = 17$. Usually more complex systems of scoring are used to produce meaningful alignments, and alignments are evaluated by likelihood or odds scores (Chapter 3), but an inverse relationship between similarity and distance scores for the alignment still holds.

A difficult problem encountered in aligning sequences is deciding whether or not a particular alignment is significant. Does a particular alignment score reveal similarity between two sequences, or would the score be just as easily found between two unrelated sequences (or random sequence of similar composition generated by the computer)? This problem was addressed by S. Karlin and S. Altschul (1990, 1993) and is addressed in detail in Chapter 3 (p. 96).

An analysis of scores of unrelated or random sequences revealed that the scores could frequently achieve a value much higher than expected in a normal distribution. Rather, the scores followed a distribution with a positively skewed tail, known as the extreme value distribution. This analysis provided a way to assess the probability that a score found between two sequences could also be found in an alignment of unrelated or random sequences of the same length. This discovery was particularly useful for assessing matches between a query sequence and a sequence database discussed in Chapter 7. In this case, the evaluation of a particular alignment score must take into account the number of sequence comparisons made in searching the database. Thus, if a score between a query protein sequence and a database protein sequence is achieved with a probability of 10^{-7} of being between unrelated sequences, and 80,000 sequences were compared, then the highest expected score (called the EXPECT score) is $10^{-7} \times 8 \times 10^{4} = 8 \times 10^{-3} = 0.008$. A value of 0.02–0.05 is considered significant. Even when such a score is found, the alignment must be carefully examined for shortness of the alignment, unrealistic amino acid matches, and runs of repeated amino acids, the presence of which decreases confidence in an alignment.

MULTIPLE SEQUENCE ALIGNMENT

In addition to aligning a pair of sequences, methods have been developed for aligning three or more sequences at the same time (for an early example, see Johnson and Doolittle 1986). These methods are computer-intensive and usually are based on a sequential aligning of the most-alike pairs of sequences. The programs commonly used are the GCG program PILEUP (http://www.gcg. com/) and CLUSTALW (Thompson et al. 1994) (Baylor College of Medicine, http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html). Once the alignment of a related set of molecular sequences (a family) has been produced, highly conserved regions (Gribskov et al. 1987) can be identified that may be common to that particular family and may be used to identify other members of the same family. Two matrix representations of the multiple sequence alignment called a PROFILE and a POSITION-SPECIFIC SCORING MATRIX (PSSM) are important computational tools for this purpose.

Multiple sequence alignments can also be the starting point for evolutionary modeling. Each column of aligned sequence characters is examined, and then the most probable phylogenetic relationship or tree that would give rise to the observed changes is identified.

Another form of multiple sequence alignment is to search for a pattern that a set of DNA or protein sequences has in common without first aligning the sequences (Stormo et al. 1982; Stormo and Hartzell 1989; Staden 1984, 1989; Lawrence and Reilly 1990). For proteins, these patterns may define a conserved component of a structural or functional domain. For DNA sequences, the patterns may specify the binding site for a regulatory protein in a promoter region or a processing signal in an RNA molecule. Both statistical and nonstatistical methods have been widely used for this purpose. In effect, these methods sort through the sequences trying to locate a series of adjacent characters in each of the sequences that, when aligned, provides the highest number of matches. Neural networks, hidden Markov models, and the expectation maximization and Gibbs sampling methods (Stormo et al. 1982; Lawrence et al. 1993; Krogh et al. 1994; Eddy et al. 1995) are examples of methods that are used. Explanations and examples of these methods are described in Chapter 4.

PREDICTION OF RNA SECONDARY STRUCTURE

In addition to methods for predicting protein structure, other methods for predicting RNA secondary structure on computers were also developed at an early time. If the complement of a sequence on an RNA molecule is repeated down the sequence in the opposite chemical direction, the regions may base-pair and form a hairpin structure, as illustrated in Figure 1.5.