CHAPTER 1

Historical Introduction and Overview

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.

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 growth of the number of sequences in GenBank can be tracked at http://www.ncbi.nlm.nih.gov/GenBank/genebankstats.html.

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.

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).

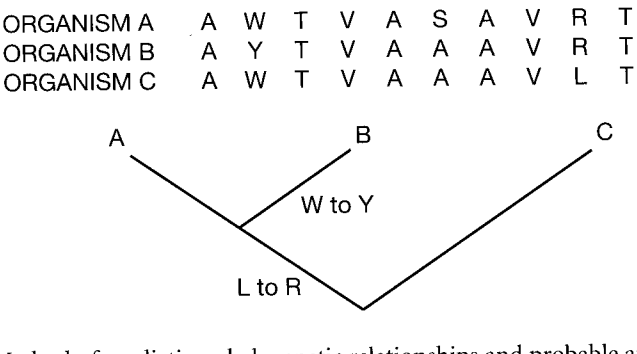
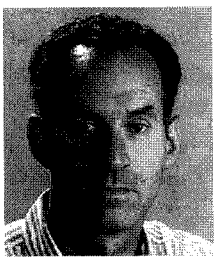


Figure 1.1. Method of predicting phylogenetic relationships and probable amino acid changes during the evolution of related protein sequences. Shown are three highly conserved sequences (A, B, and C) of the same protein from three different organisms. The sequences are so similar that each position should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

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).

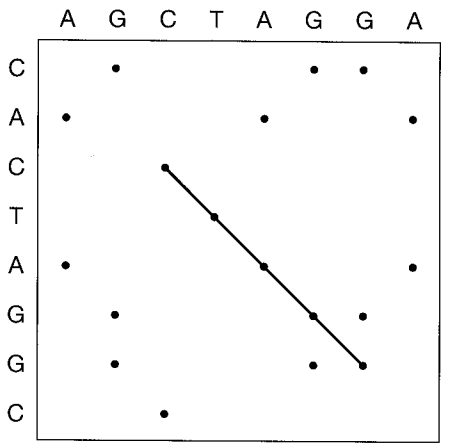


Figure 1.2. A simple dot matrix comparison of two DNA sequences, AGCTAGGA and GACTAGGC. The diagonal of dots reveals a run of similar sequence CTAGG in the two sequences.

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.

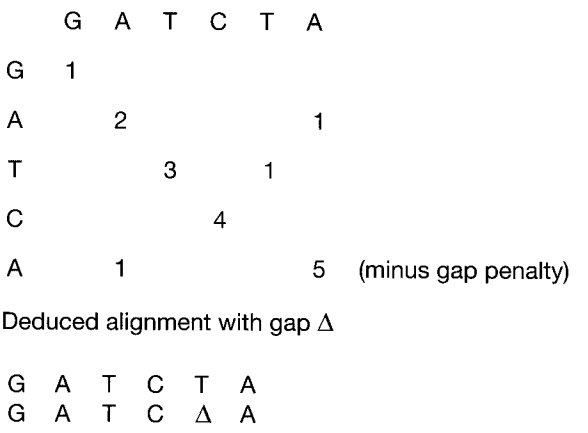
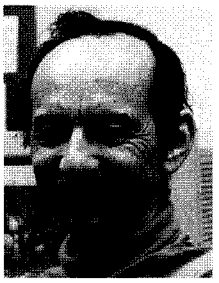


Figure 1.4. Simplified example of Needleman-Wunsch alignment of sequences GATCTA and GATCA. First, all matches in the two sequences are given a score of 1, and mismatches a score of 0 (not shown), chosen arbitrarily for this example. Second, the diagonal 1s are added sequentially, in this case to a total 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.

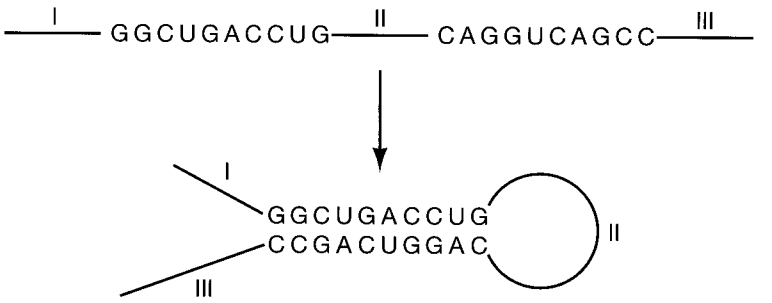


Figure 1.5. Folding of single-stranded RNA molecule into a hairpin secondary structure. Shown are portions of the sequence that are complementary: They can base-pair to form a double-stranded region. G/C base pairs are the most energetic due to 3 H bonds; A/U and G/U are next most energetic with two and one H bonds, respectively.

Tinoco et al. (1971) generated these symmetrical regions in small oligonucleotide molecules and tried to predict their stability based on estimates of the free energy associated with stacked base pairs in the model and of the destabilizing effects of loops, using a table of energy values (Tinoco et al. 1971; Salser 1978). Single-stranded loops and other unpaired regions decreased the predicted energy. Subsequently, Nussinov and Jacobson (1980) devised a fast computer method for predicting an RNA molecule with the highest possible number of base pairs based on the same dynamic programming algorithm used for aligning sequences. This method was improved by Zuker and Stiegler (1981), who added molecular constraints and thermodynamic information to predict the most energetically stable structure.

Another important use of RNA structure modeling is in the construction of databases of RNA molecules. One of the most significant of these is the ribosomal RNA database prepared by the laboratory of C. Woese (1987) (http://www.cme.msu.edu/RDP html/index.html). RNA secondary structure prediction is discussed in Chapter 5. Alignment, structural modeling, and phylogenetic analysis based on these RNA sequences have made possible the discovery of evolutionary relationships among organisms that would not have been possible otherwise.

DISCOVERY OF EVOLUTIONARY RELATIONSHIPS USING SEQUENCES

Variations within a family of related nucleic acid or protein sequences provide an invaluable source of information for evolutionary biology. With the wealth of sequence information becoming available, it is possible to track ancient genes, such as ribosomal RNA and some proteins, back through the tree of life and to discover new organisms based on their sequence (Barns et al. 1996). Diverse genes may follow different evolutionary histories, reflecting transfers of genetic material between species. Other types of phylogenetic analyses can be used to identify genes within a family that are related by evolutionary descent, called orthologs. Gene duplication events create two copies of a gene, called paralogs, and many such events can create a family of genes, each with a slightly altered, or possibly new, function. Once alignments have been produced and alignment scores found, the most closely related sequence pairs become apparent and may be placed in the outer branches of an evolutionary tree, as shown for sequences A and B in Figure 1.1 (p. 2). The next most-alike sequence, sequence C in Figure 1.1, will be represented by the next branch down on the tree. Continuing this process generates a predicted pattern of evolution for

that particular gene. Once a tree has been found, the sequence changes that have taken place in the tree branches can be inferred.

The starting point for making a phylogenetic tree is a sequence alignment. For each pair of sequences, the sequence similarity score gives an indication as to which sequences are most closely related. A tree that best accounts for the numbers of changes (distances) between the sequences (Fitch and Margoliash 1987) of these scores may then be derived. The method most commonly used for this purpose is the neighbor-joining method (Saitou and Nei 1987) described in Chapter 6. Alternatively, if a reliable multiple sequence alignment is available, the tree that is most consistent with the observed variation found in each column of the sequence alignment may be used. The tree that imposes the minimum number of changes (the maximum parsimony tree) is the one chosen (Felsenstein 1988).

In making phylogenetic predictions, one must consider the possibility that several trees may give almost the same results. Tests of significance have therefore been derived to determine how well the sequence variation supports the existence of a particular tree branch (Felsenstein 1988). These developments are also discussed in Chapter 6.

IMPORTANCE OF DATABASE SEARCHES FOR SIMILAR SEQUENCES

As DNA sequencing became a common laboratory activity, genes with an important biological function could be sequenced with the hope of learning something about the biochemical nature of the gene product. An example was the retrovirus-encoded v-*sis* and v-*src* oncogenes, genes that cause cancer in animals. By comparing the predicted sequences of the viral products with all of the known protein sequences at the time, R. Doolittle and colleagues (1983) and W. Barker and M. Dayhoff (1982) both made the startling discovery that these genes appeared to be derived from cellular genes. The Sis protein had a sequence very similar to that of the platelet-derived growth factor (PDGF) from mammalian cells, and Src to the catalytic chain of mammalian cAMP-dependent kinases. Thus, it appeared likely that the retrovirus had acquired the gene from the host cell as some kind of genetic exchange event and then had produced a mutant form of the protein that could compromise the function of the normal protein when the virus infected another animal. Subsequently, as molecular biologists analyzed more and more gene sequences, they discovered that many organisms share similar genes that can be identified by their sequence similarity.

These searches have been greatly facilitated by having genetic and biochemical information from model organisms, such as the bacterium *Escherichia coli* and the budding yeast *Saccharomyces cerevisiae*. In these organisms, extensive genetic analysis has revealed the function of genes, and the sequences of these genes have also been determined. Finding a gene in a new organism (e.g., a crop plant) with a sequence similar to a model organism gene (e.g., yeast) provides a prediction that the new gene has the same function as in the model organism. Such searches are becoming quite commonplace and are greatly facilitated by programs such as FASTA (Pearson and Lipman 1988) and BLAST (Altschul et al. 1990).

The methods used by BLAST and other additional powerful methods to perform sequence similarity searching are described further in the next section and in Chapter 7.

THE FASTA AND BLAST METHODS FOR DATABASE SEARCHES

As the number of new sequences collected in the laboratory increased, there was also an increased need for computer programs that provided a way to compare these new sequences sequentially to each sequence in the existing database of sequences, as was done

PORTION OF SEQUENCE A	–	–	W	I	V	–	–
PORTION OF SEQUENCE B	–	–	W	I	V	–	–

Figure 1.6. Rapid identification of sequence similarity by FASTA and BLAST. FASTA looks for short regions in these two amino acid sequences that match and then tries to extend the alignment to the right and left. In this case, the program found by a quick and simple indexing method that W, I, and then V occurred in the same order in both sequences, providing a good starting point for an alignment. BLAST works similarly, but only examines matched patterns of length 3 of the more significant amino acid substitutions that are expected to align less frequently by chance alone.

Bill Pearson

to identify successfully the function of viral oncogenes. The dynamic programming method of Needleman and Wunsch would not work because it was much too slow for the computers of the time; today, however, with much faster computers available, this method can be used. W. Pearson and D. Lipman (1988) developed a program called FASTA, which performed a database scan for similarity in a short enough time to make such scans routinely possible. FASTA provides a rapid way to find short stretches of similar sequence between a new sequence and any sequence in a database. Each sequence is broken down into short words a few sequence characters long, and these words are organized into a table indicating where they are in the sequence. If one or more words are present in both sequences, and especially if several words can be joined, the sequences must be similar in those regions. Pearson (1990, 1996) has continued to improve the FASTA method for similarity searches in sequence databases.

An even faster program for similarity searching in sequence databases, called BLAST, was developed by S. Altschul et al. (1990). This method is widely used from the Web site of the National Center for Biotechnology Information at the National Library of Medicine in Washington, DC (http://www.ncbi.nlm.nih.gov/BLAST). The BLAST server is probably the most widely used sequence analysis facility in the world and provides similarity searching to all currently available sequences. Like FASTA, BLAST prepares a table of short sequence words in each sequence, but it also determines which of these words are most significant such that they are a good indicator of similarity in two sequences, and then confines the search to these words (and related ones), as described in Figure 1.6. There are versions of BLAST for searching nucleic acid and protein databases, which can be used to translate DNA sequences prior to comparing them to protein sequence databases (Altschul et al. 1997). Recent improvements in BLAST include GAPPED-BLAST, which is threefold faster than the original BLAST, but which appears to find as many matches in databases, and PSI-BLAST (position-specific-iterated BLAST), which can find more distant matches to a test protein sequence by repeatedly searching for additional sequences that match an alignment of the query and initially matched sequences. These methods are discussed in Chapter 7.

PREDICTING THE SEQUENCE OF A PROTEIN BY TRANSLATION OF DNA SEQUENCES

Protein sequences are predicted by translating DNA sequences that are cDNA copies of mRNA sequences from a predicted start and end of an open reading frame. Unfortunately, cDNA sequences are much less prevalent than genomic sequences in the databases. Partial sequence (expressed sequence tags, or ESTs) libraries for many organisms are available, but these only provide a fraction of the carboxy-terminal end of the protein sequence and usually only have about 99% accuracy. For organisms that have few or no introns in their genomic DNA (such as bacterial genomes), the genomic DNA may be translated. For most

eukaryotic organisms with introns in their genes, the protein-encoding exons must be predicted and then translated by methods described in Chapter 8. These genome-based predictions are not always accurate, and thus it remains important to have cDNA sequences of protein-encoding genes. Promoter sequences in genomes may also be analyzed for common patterns that reflect common regulatory features. These types of analyses require sophisticated approaches that are also discussed in Chapter 8 (Hertz et al. 1990).

PREDICTING PROTEIN SECONDARY STRUCTURE

There are a large number of proteins whose sequences are known, but very few whose structures have been solved. Solving protein structures involves the time-consuming and highly specialized procedures of X-ray crystallography and nuclear magnetic resonance (NMR). Consequently, there is much interest in trying to predict the structure of a protein, given its sequence. Proteins are synthesized as linear chains of amino acids; they then form secondary structures along the chain, such as α helices, as a result of interactions between side chains of nearby amino acids. The region of the molecule with these secondary structures then folds back and forth on itself to form tertiary structures that include α helices, β sheets comprising interacting β strands, and loops (Fig. 1.7). This folding often leaves amino acids with hydrophobic side chains facing into the interior of the folded molecule and polar amino acids that can interact with water and the molecular environment facing outside in loops. The amino acid sequence of the protein directs the folding pathway, sometimes assisted by proteins called chaperonins. Chou and Fasman (1978) and Garnier et al. (1978) searched the small structural database of proteins for the amino acids associated with each of the secondary structure types—α helices, turns, and β strands. Sequences of proteins whose structures were not known were then scanned to determine whether the amino acids in each region were those often associated with one type of structure. For example, the amino acid proline is not often found in α helices because its side chain is not compatible with forming a helix. This method predicted the structure of some proteins well but, in general, was about as likely to predict a correct as an incorrect structure.

As more protein structures were solved experimentally, computational methods were used to find those that had a similar structural fold (the same arrangement of secondary structures connected by similar loops). These methods led to the discovery that as new protein structures were being solved, they often had a structural fold that was already known in a group of sequences. Thus, proteins are found to have a limited number of ~500 folds (Chothia 1992), perhaps due to chemical restraints on protein folding or to the exis-

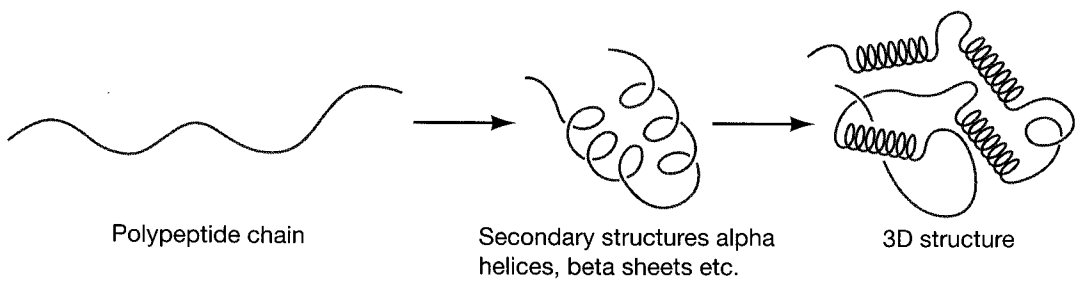


Figure 1.7. Folding of a protein from a linear chain of amino acids to a three-dimensional structure. The folding pathway involves amino acid interactions. Many different amino acid patterns are found in the same types of folds, thus making structure prediction from amino acid sequence a difficult undertaking.

tence of a single evolutionary pathway for protein structure (Gibrat et al. 1996). Furthermore, proteins without any sequence similarity could adopt the same fold, thus greatly complicating the prediction of structure from sequence. Methods for finding whether or not a given protein sequence can occupy the same three-dimensional conformation as another based on the properties of the amino acids have been devised (Bowie et al. 1991). Databases of structural families of proteins are available on the Web and are described in Chapter 9.

Amos Bairoch (Bairoch et al. 1997) developed another method for predicting the biochemical activity of an unknown protein, given its sequence. He collected sequences of proteins that had a common biochemical activity, for example an ATP-binding site, and deduced the pattern of amino acids that was responsible for that activity, allowing for some variability. These patterns were collected into the PROSITE database (http://www.expasy.ch/prosite). Unknown sequences were scanned for the same patterns. Subsequently, Steve and Jorga Henikoff (Henikoff and Henikoff 1992) examined alignments of the protein sequences that make up each MOTIF and discovered additional patterns in the aligned sequences called BLOCKS (see http://www.blocks.fhcrc.org/). These patterns offered an expanded ability to determine whether or not an unknown protein possessed a particular biochemical activity. The changes that were in each column of these aligned patterns were counted and a new set of amino acid substitution matrices, called BLOSUM matrices, similar to the PAM matrices of Margaret Dayhoff, were produced. One of these matrices, BLOSUM62, is most often used for aligning protein sequences and searching databases for similar sequences (Henikoff and Henikoff 1992) (see Chapter 7).

Sophisticated statistical and machine-training techniques have been used in more recent protein structure prediction programs, and the success rate has increased. A recent advance in this now active field of research is to organize proteins into groups or families on the basis of sequence similarity, and to find consensus patterns of amino acid domains characteristic of these families using the statistical methods described in Chapters 4 and 9. There are many publicly accessible Web sites described in Chapter 9 that provide the latest methods for identifying proteins and predicting their structures.

THE FIRST COMPLETE GENOME SEQUENCE

Although many viruses had already been sequenced, the first planned attempt to sequence a free-living organism was by Fred Blattner and colleagues (Blattner et al. 1997) using the bacterium *E. coli*. However, there was some concern over whether such a large sequence, about 4×10^6 bp, could be obtained by the then-current sequencing technology. The first published genome sequence was that of the single, circular chromosome of another bacterium, *Hemophilus influenzae* (Fleischmann et al. 1995), by The Institute of Genetics Research (TIGR, at http://www.tigr.org/), which had been started by researcher Craig Venter. The project was assisted by microbiologist Hamilton Smith, who had worked with this organism for many years. The speedup in sequencing involved using automated reading of DNA sequencing gels through dye-labeling of bases, and breaking down the chromosome into random fragments and sequencing these fragments as rapidly as possible without knowledge of their location in the whole chromosome. Computer analysis of such shotgun cloning and sequencing techniques had been developed much earlier by R. Staden at Cambridge University and other workers, but the TIGR undertaking was much more ambitious. In this genome project, newly read sequences were immediately entered into a computer database and compared with each other to find overlaps and produce contigs of two or more sequences with the assistance of computer programs. This procedure circumvented the need to grow and keep track of large numbers of subclones. Although the same

sequence was often obtained up to 10 times, the sequence of the entire chromosome (2×10^9 bp), less a few gaps, was rapidly assembled in the computer over a 9-month period at a cost of about $\$10^6$.

This success heralded a large number of other sequencing projects of various prokaryotic and eukaryotic microorganisms, with a tremendous potential payoff in terms of utilizable gene products and evolutionary information about these organisms. To date, completed projects include more than 30 prokaryotes, yeast *S. cerevisiae* (see Cherry et al. 1997), the nematode *Caenorhabditis elegans* (see *C. elegans* Sequencing Consortium 1998), and the fruit fly *Drosophila* (see Adams et al. 2000). The plant *Arabidopsis thaliana* and the human genome sequencing projects are ongoing and will be completed during 2000 or shortly thereafter.

The Human Genome Project, a large, federally funded collaborative project, will complete sequencing of the entire human genome by 2003. The project was developed from an idea discussed at scientific meetings in 1984 and 1985, and a pilot project, the Human Genome Initiative, was begun by the Department of Energy (DOE) in 1986. National Institutes of Health funding of the project began in 1987 under the Office of Genome Research. Currently, the project is constituted as the National Human Genome Research Initiative. In 1998, a new commercial venture under the leadership of Craig Venter was formed to sequence the majority of the human genome by 2001. This group, which uses a whole genome shotgun cloning approach and intensive computer processing of data, has already completed the Drosophila sequence and will sequence the mouse genome following completion of the human genome. Both groups simultaneously announced completion of the sequencing of the human genome in 2000.

ACEDB, THE FIRST GENOME DATABASE

As more genetic and sequence information became available for the model organisms, interest arose in generating specific genome databases that could be queried to retrieve this information. Such an enterprise required a new level of sharing of data and resources between laboratories. Although there were initial concerns about copyright issues, credits, accuracy, editorial review, and curating, eventually these concerns disappeared or became resolved as resources on the Internet developed. The first genome database, called ACEDB (a *C. elegans* database), and the methods to access this database were developed by Mike Cherry and colleagues (Cherry and Cartinhour 1993). This database was accessible through the internet and allowed retrieval of sequences, information about genes and mutants, investigator addresses, and references. Similar databases were subsequently developed using the same methods for *A. thaliana* and *S. cerevisiae*. Presently, there is a large number of such publicly available databases. Web access to these databases is discussed in Chapter 10 (Table 10.1, p. 482).

REFERENCES

Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bairoch A., Bucher P., and Hofmann K. 1997. The PROSITE database, its status in 1997. *Nucleic Acids Res.* **25:** 217–221.

Barker W.C. and Dayhoff M.O. 1982. Viral *src* gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase. *Proc. Natl. Acad. Sci.* **79:** 2836–2839.

Barns S.M., Delwiche C.F., Palmer J.D., and Pace N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci.* **93:** 9188–9193.

Blattner F.R., Plunkett III, G., Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., and Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

Bowie J.U., Luthy R., and Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253:** 164–170.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans:* A platform for investigating biology. *Science* **282:** 2012–2018.

Cherry J.M. and Cartinhour S.W. 1993. ACEDB, a tool for biological information. In *Automated DNA sequencing and analysis* (ed. M. Adams et al.). Academic Press, New York.

Cherry J.M., Ball C., Weng S., Juvik G., Schmidt R., Adler C., Dunn B., Dwight S., Riles L., Mortimer R. K., and Botstein D. 1997. Genetic and physical maps of *Saccharomyces cerevisiae. Nature* (suppl. 6632) **387:** 67–73.

Chothia C. 1992. Proteins. One thousand families for the molecular biologist. *Nature* **357:** 543–544.

Chou P.Y. and Fasman G.D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47:** 45–147.

Dayhoff M.O., Ed. 1972. *Atlas of protein sequence and structure*, vol. 5. National Biomedical Research Foundation, Georgetown University, Washington, D.C.

———. 1978. Survey of new data and computer methods of analysis. In *Atlas of protein sequence and structure*, vol. 5, suppl. 3. National Biomedical Research Foundation, Georgetown University, Washington, D.C.

Doolittle R.F., Hunkapiller M.W., Hood L.E., Devare S.G., Robbins K.C., Aaronson S.A., and Antoniades H.N. 1983. Simian sarcoma *onc* gene v-*sis* is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science* **221:** 275–277.

Eddy S.R., Mitchison G., and Durbin R. 1995. Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.* **2:** 9–23.

Ewing B. and Green P. 1998. Base-calling of automated sequence traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing B., Hillier L., Wendl, M.C., and Green P. 1998. Base-calling of automated sequence traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Felsenstein J. 1988. Phylogenies from molecular sequences: Inferences and reliability. *Annu. Rev. Genet.* **22:** 521–565.

Fitch W.M. and Margoliash E. 1987. Construction of phylogenetic trees. *Science* **155:** 279–284.

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Garnier J., Osguthorpe D.J., and Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120:** 97–120.

Gibbs A.J. and McIntyre G.A. 1970. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* **16:** 1–11.

Gibrat J.F., Madej T., and Bryant S.H. 1996. Surprising similarity in structure comparison. *Curr. Opin. Struct. Biol.* **6:** 377–385.

Gribskov M., McLachlan A.D., and Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci.* **84:** 4355–4358.

Henikoff S. and Henikoff J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Hertz G.Z., Hartzell III, G.W., and Stormo G.D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **6:** 81–92.

Johnson M.S. and Doolittle R.F. 1986. A method for the simultaneous alignment of three or more amino acid sequences. *J. Mol. Evol.* **23:** 267–268.

Karlin S. and Altschul S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.* **87:** 2264–2268.

———. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci.* **90:** 5873–5877.

Krogh A., Brown M., Mian I.S., Sjölander K., and Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235:** 1501–1531.

Lawrence C.E. and Reilly A.A. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins Struct. Funct. Genet.* **7:** 41–51.

Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., and Wootton J.C. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262:** 208–214.

Maizel Jr., J.V. and Lenk R.P. 1981. Enhanced graphic matrix analyses of nucleic acid and protein synthesis. *Proc. Natl. Acad. Sci.* **78:** 7665–7669.

Maxam A.M. and Gilbert W. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* **74:** 560–564.

Needleman S.B. and Wunsch C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Nussinov R. and Jacobson A.B. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci.* **77:** 6903–6913.

Pearson W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

———. 1996. Effective protein sequence comparison. *Methods Enzymol.* **266:** 227–258.

Pearson W.R. and Lipman D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Saitou N. and Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Salser W. 1978. Globin mRNA sequences: Analysis of base pairing and evolutionary implications. *Cold Spring Harbor Symp. Quant. Biol.* **42:** 985–1002.

Sanger F. and Tuppy H. 1951. The amino acid sequence of the phenylalanyl chain of insulin. *Biochem. J.* **49:** 481–490.

Sanger F., Nicklen S., and Coulson A.R. 1977. DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463–5467.

Smith T.F. and Waterman M.S. 1981a. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

———. 1981b. Comparison of biosequences. *Adv. Appl. Math.* **2:** 482–489.

Staden R. 1984. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* **12:** 505–519.

———. 1989. Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.* **5:** 89–96.

Stormo G.D. and Hartzell III, G.W. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci.* **86:** 1183–1187.

Stormo G.D., Schneider T.D., Gold L., and Ehrenfeucht A. 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10:** 2997–3011.

Thompson J.D., Higgins D.G., and Gibson T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Tinoco Jr., I., Uhlenbeck O.C., and Levine M.D. 1971. Estimation of secondary structure in ribonucleic acids. *Nature* **230:** 362–367.

Waterman M.S., Ed. 1989. Sequence alignments. In *Mathematical methods for DNA sequences*. CRC Press, Boca Raton, Florida.

Woese C.R. 1987. Bacterial evolution. *Microbiol. Rev.* **51:** 221–271.

Zuker M. and Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148.

Additional Reading

Reference Books and Special Journal Editions

Baldi P. and Brunck S. 1998. *Bioinformatics: The machine learning approach.* MIT Press, Cambridge, Massachusetts.

Baxevanis A.D. and Ouellette B.F., Eds. 1998. *Bioinformatics: A practical guide to the analysis of genes and proteins.* John Wiley & Sons, New York.

Doolittle R.F. 1986. *Of URFS and ORFS: A primer on how to analyze derived amino acid sequences.* University Science Books, Mill Valley, California.

———, Ed. 1990. Molecular evolution: Computer analysis of protein and nucleic acid sequences. *Methods Enzymol.*, vol. 183. Academic Press, San Diego.

———, Ed. 1996. Computer methods for macromolecular sequence analysis. *Methods Enzymol.*, vol. 266. Academic Press, San Diego, California.

Durbin R., Eddy S., Krogh A., and Mitchison G., Eds. 1998. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, United Kingdom.

Gribskov M. and Devereux J., Eds. 1991. *Sequence analysis primer.* University of Wisconsin Biotechnology Center Biotechnical Resource Ser. (ser. ed. R.R. Burgess). Stockton Press, New York.

Gusfield D. 1997. *Algorithms on strings, trees, and sequences: Computer science and computational biology.* Cambridge University Press, Cambridge, United Kingdom.

Martinez H., Ed. 1984. Mathematical and computational problems in the analysis of molecular sequences (special commemorative issue honoring Margaret Oakley Dayhoff). *Bull. Math. Biol.* Pergamon Press, New York.

Nucleic Acids Research. 1996–2000. Special database issues published in the January issues of volumes 22–26. Oxford University Press, Oxford, United Kingdom.

Salzberg S.L., Searls D.B., and Kasif S., Eds. 1999. Computational methods in molecular biology. *New Compr. Biochem.*, vol. 32. Elsevier, Amsterdam, The Netherlands.

Sankoff D. and Kruskal J.R., Eds. 1983. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison.* Addison-Wesley, Don Mills, Ontario.

Söll D. and Roberts R.J., Eds. 1982. The application of computers to research on nucleic acids. I. *Nucleic Acids Res.*, vol. 10. Oxford University Press, Oxford, United Kingdom.

———. 1984. The application of computers to research on nucleic acids. II. *Nucleic Acids Res.*, vol. 12. Oxford University Press, Oxford, United Kingdom.

von Heijne G. 1987. *Sequence analysis in molecular biology — Treasure trove or trivial pursuit.* Academic Press, San Diego, California.

Waterman M.S., Ed. 1989. Mathematical analysis of molecular sequences (special issue). *Bull. Math. Biol.* Pergamon Press, New York.

———. 1995. *Introduction to computational biology: Maps, sequences, and genomes.* Chapman and Hall, London, United Kingdom.

Yap, T.K., Frieder O., and Martino R.L. 1996. *High performance computational methods for biological sequence analysis.* Kluwer Academic, Norwell, Massachusetts.

Journals That Routinely Publish Papers on Sequence Analysis

Bioinformatics (formerly *Comput. Appl. Biosci.* [CABIOS]). Oxford University Press, Oxford, United Kingdom. http://bioinformatics.oupjournals.org/cabios/.

Journal of Computational Biology. Mary Ann Liebert, Larchmont, New York. http://www-hto.usc.edu/jcb/.

Journal of Molecular Biology. Academic Press, London, United Kingdom. http://www.hbuk.co.uk/jmb.

Nucleic Acids Research (sections on Genomics and Computational Biology). Oxford University Press, Oxford, United Kingdom. http://nar.oupjournals.org.