

# Sequence Alignment

Julie D Thompson, *Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France*

Olivier Poch, *Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France*

Definition sequence comparison or alignment is the cornerstone of bioinformatics, providing the basis for sequence database searching, three-dimensional structure modeling and evolutionary studies. A sequence alignment shows how a set of sequences may be related by identifying and arranging in columns the structurally and functionally equivalent residues common to all the sequences.

## Introduction

Sequence comparison or alignment plays a fundamental role in most areas of modern molecular biology, from shaping our basic conceptions of life and its evolutionary processes, to providing the foundation for the new biotechnology industry. The currently accepted universal tree of life, in which the living world is divided into three domains (Bacteria, Archaea and Eukarya), was constructed from comparative analyses of ribosomal RNA sequences. However, this view, in which life evolved from simple prokaryotes to eukaryotes, has been somewhat shaken by studies of the complete sequences of several microbial genomes, which discovered widespread evidence of significant horizontal gene transfers between diverse organisms. Comparisons of the complete sequences of the more closely related genomes have revealed that far from being static, genomic DNA is actually highly plastic, being subject to recombinations, rearrangements, insertions and deletions. (*See Comparative Genomics; DNA Recombination; Ribosomes and Ribosomal Proteins.*)

In addition to comparing whole-genome sequences, the genome sequencing projects are also attempting to determine and identify the individual genes encoded by the genome. Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes available in the sequence databases, in order to identify potential homologs, that is, sequences that have evolved from a common ancestor. Generally, homologous proteins share the same three-dimensional (3D) structure and have similar functions, active sites or binding domains. In most genome annotation projects, the standard strategy to determine the function of a novel protein is, therefore, to search the sequence databases for homologs and to transfer the structural/functional annotation from the known to the unknown protein. (*See Gene Structure and Organization; Homologous,*

*Orthologous and Paralogous Genes; Protein Characterization in Proteomics.*)

A combination of sequential and structural information has been shown to increase the accuracy of structure prediction, both 2D and 3D, relative to the exclusive use of sequence or structure. The comparison of homologous sequences provides one of the most efficient methods of modeling the 3D structure of a protein. Protein threading techniques, that rely on matching 3D information predicted from the query sequence with corresponding features of a known structure, can also be improved by incorporating sequence information. (*See Protein Homology Modeling; Protein Structure Prediction and Databases.*)

Sequence alignments on the genomic scale are also having a widespread effect in the pharmaceutical industry, providing an opportunity to identify the proteins associated with a particular disease, which are therefore potential drug targets. Recent advances in the computational analyses of enzyme structures and functions have improved the strategies used to modify enzyme specificities and mechanisms by site-directed mutagenesis, and to engineer biocatalysts through molecular reassembly. The analysis of genomes of extremophile microorganisms has led to the identification of many enzymes showing activity and stability at extremes of temperature, pH, pressure and salinity, many of which have potential for industrial and biotechnological applications. The potential impact on protein structure prediction, biology, protein engineering and medicine is enormous. (*See Drug Metabolic Enzymes; Genetic Polymorphisms; Gene Targeting by Homologous Recombination; Protein Targeting.*)

## What is an Alignment?

There exist two main categories of sequence alignment: pairwise alignment (or the alignment of two sequences) and multiple alignment. Pairwise alignments are most

### Advanced article

#### Article contents

- Introduction
- What is an Alignment?
- What is the Best Alignment?
- Alignment Algorithms

doi: 10.1038/npg.els.0005318

commonly used in database search programs such as BLAST and FASTA in order to detect homologs of a novel sequence. Multiple alignments, containing from three to several hundred sequences, are more computationally complex than pairwise alignments and, in general, simultaneous alignment of more than a few sequences is rarely attempted. Instead a series of pairwise alignments are performed and amalgamated into a multiple alignment. Nevertheless, multiple alignments have the advantage of providing an overall view of the family, thus helping to decipher the evolutionary history of the protein family. Multiple sequence alignments are useful in identifying conserved patterns in protein families, which may not be evident from pairwise alignments. They are also used in the determination of domain organization, to help predict protein secondary/tertiary structure and in phylogenetic studies. (See Multiple Alignment; Similarity Search.)

The purpose of any sequence alignment, whether pairwise or multiple, is to show how a set of sequences may be related, in terms of conserved residues, substitutions, insertion–deletion events (‘indels’). In the most general terms, an alignment represents a set of sequences using a single-letter code for each amino acid (for protein sequences) or nucleotide (for DNA/RNA sequences). Structurally and functionally equivalent residues are aligned either in rows or, more usually in columns (Figure 1). When the sequences are of different lengths, insertion–deletion events are postulated to explain the variation, and gap characters are introduced into the alignment. Sequence alignments can be further divided into global alignments that align the complete sequences and local alignments that identify only the most similar segments or sequence patterns (motifs). While global alignment algorithms produce more accurate alignments for proteins of similar length, local alignment algorithms are better at identifying similar regions within sequences when the sequences are not related over their entire length. (See Global Alignment; Sequence Similarity.)

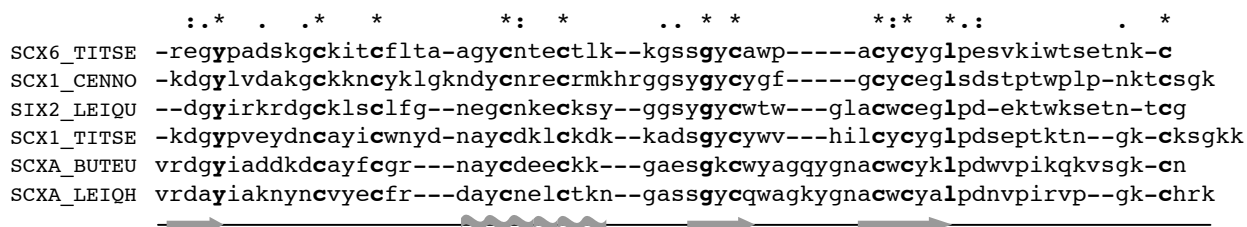
## What is the Best Alignment?

For any two sequences, there are an exponential number of potential alignments with gaps. Therefore, it is critical to be able to distinguish ‘good’ alignments from ‘bad’ ones. A good alignment is one that corresponds to the biologically correct alignment, accurately reflecting the evolutionary, structural and functional relationships between the sequences. Sequence alignment programs have, until recently, used only the primary sequence information to reconstruct these complex relationships. In order to find the best alignment, most alignment programs assign a similarity score to all possible alignments and try to maximize this score. These alignment scores, also known as objective functions, are generally based on scores for aligning single residues with penalties for introducing gaps into the sequences. While these scores are generally adequate for the alignment of relatively well-conserved sequences, it is clear that more elaborate scoring schemes will be required for the highly complex proteins detected by today’s advanced database searching methods.

## Scoring matrices

Most alignment programs make comparisons between pairs of bases or amino acids by looking up a value in a scoring matrix. The matrix contains a score for the match quality of every possible pair of residues (Figure 2). The simplest way to score an alignment is to count the number of identical residues that are aligned. When the sequences to be aligned are closely related, this will usually find approximately the correct solution. For more divergent sequences sharing less than 25–30% identity, however, the scores given to nonidentical residues become critically important.

More sophisticated scoring schemes exist for both DNA and protein sequences and generally take the form of a matrix defining the score for aligning each pair of residues. For alignments of nucleotide sequences, the simplest scoring matrix would assign



**Figure 1** Alignment of six scorpion toxin proteins. Conserved positions are shown in bold. Gaps are represented by dashes between the letter strings. The secondary structure elements of the scorpion *Leirus quinquestriatus hebraeus* protein (SCXA\_LEIQH) are shown below the alignment. Right arrow:  $\beta$ -sheet; coil:  $\alpha$ -helix.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-3	1	1	1	-6	-3	0	0	0	0	0	
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	-2	-4	-2	-1	0	-1	
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	0	-1	-1	-3	0	-2	1	2	1	-1	
I	-1	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-1	
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	
P	1	0	0	-1	-3	0	-1	0	0	-1	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	0	-1
S	1	0	1	0	0	-1	0	1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	0	
T	-1	-1	0	0	-2	-1	0	0	1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	
W	-6	-2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	0	10	-2	-3	-4	-2	-2	
V	0	-2	-2	-2	-2	-1	-2	4	-2	2	-2	2	-1	-1	1	0	0	-6	-2	4	-2	-2	-1	
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	-5	-3	-2	3	2	2	-1	
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	

**Figure 2** PAM250 matrix. Substitution scores for amino acids.

the same score to a match of the four classes of bases, ACGT, and 0 for any mismatch. However, transitions (substitution of A–G or C–T) happen much more frequently than transversions (substitution of A–T or G–C), and it is often desirable to score these substitutions differently. More complex matrices also exist in which matches between ambiguous nucleotides are given values whenever there is any overlap in the sets of nucleotides represented by the two symbols being compared. For protein sequence comparisons, scoring matrices generally take into account the biochemical similarities between residues and/or the relative frequencies with which each amino acid is substituted by another.

The most widely used scoring matrices are known as the PAM (point accepted mutation) matrices (Dayhoff *et al.*, 1978). The original PAM1 matrix construct was based on the mutations observed in a large number of alignments of closely related sequences. A series of matrices was then extrapolated from the PAM1. The matrices range from strict ones, useful for comparing very closely related sequences, to very ‘soft’ ones that are used to compare very divergent sequences. For example, the PAM250 matrix corresponds to an evolutionary distance of 250%, or approximately 80% residue divergence. Other matrices have been derived directly from either sequence-based or structure-based alignments, such as the Blosum matrices, which are based on the observed residue substitutions in aligned sequence segments from the Blocks database. The proteins in the database are clustered at different percentage identities to produce a series of matrices. For example, the Blosum-62 matrix is based on alignment blocks in which all the sequences share at least 62% residue identity. Other more specialized matrices have been developed, for example, for specific secondary structure elements (helices,  $\beta$ -sheets), or for the comparison of particular types of proteins such as transmembrane proteins. (See Substitution Matrices.)

## Gap schemes

In addition to assigning scores for residue matches and mismatches, most alignment scoring schemes in use today calculate a cost for the insertion of gaps in the sequences. One of the first gap scoring schemes for the alignment of two sequences charged a fixed penalty for each residue in either sequence aligned with a gap in the other. Under this system, the cost of a gap is proportional to its length. Alignment algorithms implementing such length-proportional-gap penalties are efficient; however, the resulting alignments often contain a large number of short indels that are not biologically meaningful. To address this problem, linear or ‘affine’ gap costs are used that define a gap insertion or ‘gap opening’ penalty in addition to the length-dependent or ‘gap extension’ penalty. Thus, a smaller number of long gaps is favored over many short ones. Fortunately, algorithms using affine gap costs are only slightly more complex than those using length-proportional gap penalties, requiring only a constant factor more space and time.

Again, more complex schemes have been developed, such as ‘concave’ gap costs or position-specific gap penalties. Most of these are attempts to mimic the biological processes or constraints that are thought to regulate the evolution of DNA or protein sequences. (See Exons: Insertion and Deletion during Evolution; Gross Insertions and Microinsertions in Evolution.)

## Alignment statistics

An important aspect of sequence alignment is to establish how meaningful a given alignment is. It is always possible to construct an alignment between a set of sequences, even if they are unrelated. The problem is to determine the level of similarity required to infer that the sequences are homologous. A simple rule-of-thumb for protein sequences states that if two sequences share more than 25% identity over more than 100 residues, then the two sequences can be assumed to be homologous. However, many proteins sharing less than 25% residue identity, said to be in the ‘twilight zone’ (Doolittle, 1986), do still have very similar structures. The measure of the percentage identity or similarity of the sequences is generally not sensitive enough to distinguish between alignments of related and unrelated sequences. (See Evolutionary Distance; Gene Families: Formation and Evolution; Gene Structure: Evolution; Protein Structure.)

Much work has been done on the significance of both ungapped and gapped pairwise local alignments (Altschul and Gish, 1996; Pearson, 1998), although the statistics of global alignments or alignments of more than two sequences are far less well understood. The aim of the statistical analysis is to estimate the

probability of finding by ‘chance’ at least one alignment that scores as high as or greater than the given alignment. For ungapped local alignments, these probabilities or *P*-values may be derived analytically. For alignments with gaps, empirical estimates are used, based on the scores obtained during a database search, or from randomly generated sequences.

For database search programs, the significance of an alignment between the query sequence and a database sequence is often expressed in terms of expect- or *E*-values, which specifies the number of matches with a given score that are expected to occur by chance in a search of a database. An *E*-value of zero, with a given score, would indicate that no matches with this score are expected purely by chance. (See Alignment: Statistical Significance.)

## Alignment Algorithms

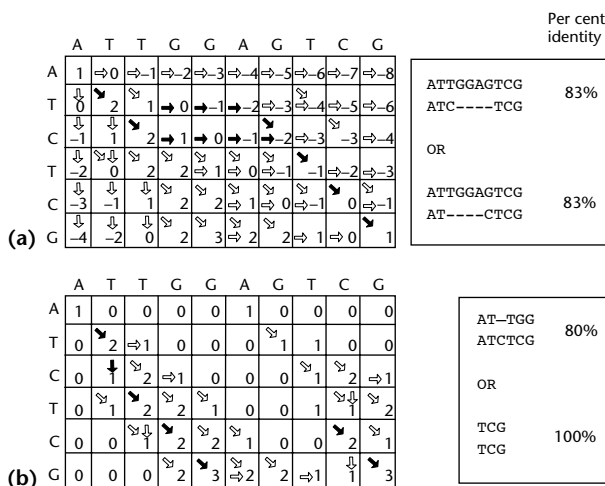
### Pairwise alignments

The comparison or alignment of biological sequences began in the early 1970s, with the first dynamic programming algorithm for the global (or full-length) alignment of two sequences (Needleman and Wunsch, 1970). The optimal local alignment between a pair of sequences, in which only the highest scoring subsegments of the two sequences are aligned, involves a simple modification to the Needleman–Wunsch method (Smith and Waterman, 1981). (See Smith–Waterman Algorithm.)

Dynamic programming is a rigorous mathematical technique that is guaranteed to find the maximal scoring alignment for any two sequences. It does this by constructing a 2D alignment matrix or path graph of partial alignment scores (Figure 3). Each position in the matrix contains the score for the best partial alignment that ends at that position. The best scoring partial alignment will be extended to subsequent positions in the matrix, either by aligning one residue from each sequence, or by inserting a gap into one or other of the sequences. In this way, all possible alignments are considered and the final alignment is thus the best-scoring alignment possible. The optimal global alignment score is given in the bottom, right-hand corner of the alignment matrix, while the optimal local alignment score is defined as the highest-scoring position anywhere in the alignment matrix. (See Dynamic Programming.)

### Heuristic methods

A different approach to the local alignment problem involves the use of heuristics or ‘approximate’ methods, which do not guarantee an optimal alignment solution



**Figure 3** Dynamic programming alignment matrices for global (a) and local (b) alignments of two DNA sequences. Per cent identity scores for each alignment are calculated by dividing the number of identical residues aligned by the total number of residues aligned.

but are less time-consuming than the rigorous dynamic programming techniques. These approximate alignment algorithms are used in programs such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990) to search the protein and DNA sequence databases for homologs of a target sequence. The general approach involves comparing the target or ‘query’ sequence to all the sequences in a specified database in a pairwise fashion. Each comparison is given a score reflecting the degree of similarity between the query and the sequence being compared. The higher the score, the greater the degree of similarity. The similarity is measured and shown by aligning the two sequences. (See BLAST Algorithm; FASTA Algorithm.)

The heuristics used involve finding patches of regional similarity, rather than trying to find the best alignment between the entire query and an entire database sequence. FASTA uses a two-step pairwise alignment algorithm. The first step consists of a search for exactly matching strings or ‘words’ that are common to both sequences. This is done in order to identify regions in a 2D table similar to that shown for the dynamic programming algorithm above, and are likely to correspond to highly similar segments shared by the two sequences. These regions will consist of a diagonal or a few closely spaced diagonals in the table, which have a high number of word matches between the sequences. The second step involves a Smith–Waterman local alignment centered on these regions. The speed-up achieved by a FASTA alignment relative to a full Smith–Waterman alignment is due to the restriction of the dynamic programming algorithm to only the high-scoring regions. In the BLAST program, the first step also involves a word-based heuristic,

similar to that of FASTA. However, the high-scoring segments found are then extended in both forward and backward directions to generate an alignment that continues until the sequence ends, or the alignment becomes nonsignificant. In both FASTA and BLAST, in addition to the alignment scores, the significance of each alignment is computed as a *P*-value or an *E*-value (see the section Alignment statistics).

## Dot plots

A dot plot is a powerful, visual method for comparing two complete sequences that provides a global view of all possible regions of similarity between the two sequences. Dot plot programs often provide an interactive environment in which the user can select significant sequence segments in order to guide the final alignment.

In the dot plot in **Figure 4**, the *X* and *Y* axes of the plot correspond to the two sequences to be compared. The dots represent all the possible matches of identical residues in the two sequences. Any region of similar sequence appears as a diagonal row of dots. Isolated dots not on the diagonal represent random matches, which are probably not related to any significant alignment.

Visualization of matching regions may be improved by filtering out these random matches using a sliding window calculation. Instead of comparing single

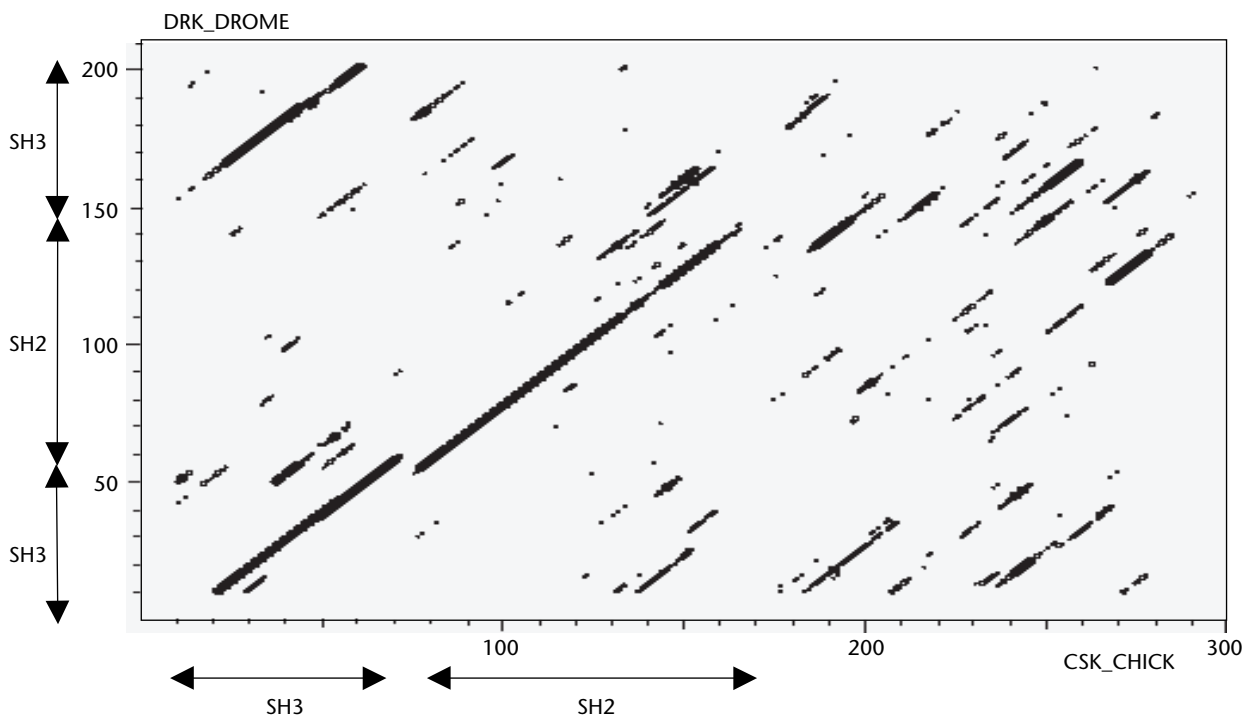
sequence positions in the two sequences, the average score in a window of adjacent positions is calculated, and a dot is printed only if the score for the window is above a certain average score. Scoring matrices such as the PAM or Blosom matrices may be used instead of residue identities.

Dot plots are particularly valuable for finding repeats or inversions in protein and DNA sequences, and for predicting regions in RNA that are self-complementary and that, therefore, might form a double-stranded region or secondary structure. For an excellent description of the dot plot method, see States and Boguski (1991).

## Multiple sequence alignment

The first formal algorithm for multiple sequence alignment (Sankoff, 1975) was developed as a direct extension of the pairwise dynamic programming algorithm. However, the optimal multiple alignment of more than a few sequences (more than 10) remains impractical due to the intensive computer resources required, despite some recent space and time improvements. Therefore, in order to multiply align larger sets of sequences, most programs in use today employ some kind of heuristic approach to reduce the problem to a reasonable size.

Traditionally, the most popular method has been the progressive alignment procedure (Feng and



**Figure 4** Dot plot of a chicken tyrosine-protein kinase protein (CSK\_CHICK) compared to a *Drosophila* SH2–SH3 adaptor protein (DRK\_DROME).

Doolittle, 1987). A multiple sequence alignment is built up gradually using a series of pairwise alignments. The two closest sequences are aligned first and then larger and larger sets of sequences are merged, until all the sequences are included in the multiple alignment. Programs implementing the progressive multiple alignment method may use either a local or a global alignment algorithm.

One of the most widely used progressive multiple alignment programs is CLUSTAL W (Thompson *et al.*, 1994). More recently, algorithms other than dynamic programming have been exploited in the search for more accurate multiple alignments in a wider variety of situations. New developments include the use of hidden markov models (HMMs), genetic algorithms, segment-to-segment alignments, Gibbs sampling or iterative refinement techniques. (See Hidden Markov Models.)

### See also

Alignment: Statistical Significance  
BLAST Algorithm  
FASTA Algorithm  
Global Alignment  
Multiple Alignment

### References

- Altschul SF and Gish W (1996) Local alignment statistics. *Methods in Enzymology* **266**: 460–480.
- Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.
- Dayhoff M, Schwartz RM and Orcutt BC (1978) A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, vol. 5(supplement 3), pp. 345–358. Silver Springs, MD: National Biomedical Research Foundation.
- Doolittle RF (1986) *Of Urfs and Orfs: Primer on How to Analyze Derived Amino Acid Sequences*. Mill Valley, CA: University Science Books.
- Feng DF and Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**: 351–360.
- Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**: 443–453.
- Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. *Journal of Molecular Biology* **276**: 71–84.
- Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 2444–2448.
- Sankoff D (1975) Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics* **78**: 35–42.
- Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **215**: 403–410.
- States DJ and Boguski MS (1991) Similarity and homology. In: Gribskov M and Devereux J (eds.) *Sequence Analysis Primer*, pp. 92–124. New York, NY: Stockton Press.
- Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and matrix choice. *Nucleic Acids Research* **22**: 4673–4680.
- Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology* **219**: 555–565.
- Apostolico A and Giancarlo R (1998) Sequence alignment in molecular biology. *Journal of Computational Biology* **5**: 173–196.
- Benner SA, Cohen MA and Gonnet GH (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of Molecular Biology* **229**: 1065–1082.
- Durbin R, Eddy S, Krogh A and Mitchison G (1999) Pairwise alignment. In: Durbin R (ed.) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, pp. 12–45. Cambridge, UK: Cambridge University Press.
- Gotoh O (1999) Multiple sequence alignment: algorithms and applications. *Advanced Biophysics* **36**: 159–206.
- Henikoff S and Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. *Proteins* **17**: 49–61.
- Henikoff S (1994) Comparative sequence analysis: finding genes. In: Smith DW (ed.) *Biocomputing, Informatics and Genome Projects*, pp. 87–117. New York, NY: Academic Press.
- Smith TF (1999) The art of matchmaking: sequence alignment methods and their structural implications. *Structure with Folding and Design* **7**: R7–R12.
- Vogt G, Etzold T and Argos P (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *Journal of Molecular Biology* **249**: 816–831.
- Waterman MS (1995) Dynamic programming alignment of two sequences. In: Michael SW (ed.) *Introduction to Computational Biology: Maps, Sequences and Genomes*, pp. 183–232. London, UK: Chapman & Hall/CRC Press.
- Yona G and Brenner SE (2000) Comparison of protein sequences and practical database searching. In: Higgins DG and Taylor WR (eds.) *Bioinformatics: Sequence, Structure and Databases. A Practical Approach*, pp. 167–190. Oxford, UK: Oxford University Press.

### Further Reading



### Web Links

- Blocks database  
<http://www.blocks.fhcrc.org/>