Pairwise Sequence Alignment Methods

1. Dot matrix analysis.

One sequence is plotted against the other sequence and matches of amino acid type are indicated by dots. Matching regions are shown as diagonal lines of dots. All regions of matching sequence between the two proteins can be found. This has the advantage of locating repeated and inverted sequences, e.g. duplicated domains, which can be missed by the other methods that give one or a small number of optimal alignments, rather than all possibilities.

2. Dynamic programming algorithm.

In this method each residue in the two sequences is compared using a scoring scheme for matches, mismatches, and gaps. A matrix is generated that represents all possible alignments of residues or characters. This matrix is searched for the alignment with the highest set of sequential scores, which will be the optimal alignment. It will always find the highest scoring alignment. But, other local alignments may be of interest, e.g. in the case of sequence repeats. This method is slow for longer sequences, and impractical for major database searching.

Dynamic programming requires a substitution matrix to score for matched and mismatched characters (or residues), and a weighting scheme to penalize gaps.

Running Best Score $$S_{ij} = \max \{ S_{i-1,j-1} + s(a_i b_j), \; \max_{x > 1)} (S_{i-x,j} - w_x), \; \max_{y > 1} (S_{i,j-y} - w_y) \}$$

where S_{ij} is the score at position i in sequence a and position j in sequence b,
$s(a_i b_j)$ is the score for aligning the characters at positions i and j,
w_x is the penalty for a gap of length x in sequence a, and
w_y is the penalty for a gap of length y in sequence b.

The final best score will be the highest value for the running best score S_{ij}.

3. Word or Ktuple-based methods.

Database searching programs for sequence alignment such as FASTA and BLAST speed up the search by first looking for matching short sequences called words or ktuples. These methods are heuristic, i.e. empirical methods are used to find solution. High scoring words are joined and extended by dynamic programming to obtain an alignment. These methods are usually statistically reliable.

FASTA uses a recommended word size of 2 (or 1 for finding distant relationships) for protein sequences, and uses 4-6 for nucleic acids. The word size can be changed.

BLAST uses a word size of 3 for proteins and 11 for nucleic acids. The BLAST search may be faster than FASTA but less sensitive due to the longer word size.

Matrix to Align Two Sequences

E**QR**I**AL**NT**L**K**DYAMR**IG**LN**T**LKD**F**--GKYQ**
D**QR**V**AL--L**R**DYAMR**FA**LN**S**LKD**YGL**GKYQ**

	E	Q	R	I	A	L	N	T	L	K	D	Y	A	M	R	I	G	L	N	T	L	K	D	F	G	K	Y	Q
D	○																											
Q		**x**					○												○									x
R			**x**							○					x							○				○		
V				○		○			○					○		○		○			○							
A					**x**								x															
L				○		**x**			x					○		○		x			x							
L				○		x			**x**					○		○		x			x							
R			x							○					x							○				○		
D	○										**x**												x					
Y												**x**												○			x	
A					x								**x**															
M				○		○			○					**x**		○		○			○							
R			x							○					**x**								○			○		
F												○												x			○	
A					x								x															
L				○		x			x					○		○		**x**			x							
N		○					x												**x**									○
S								○												○								
L				○		x			x					○		○		x			**x**							
K			○							x					○							**x**				x		
D	○										x												**x**					
Y												x												○			x	
G																	x								x			
L				○		x			x					○		○		x			x							
G																	x								**x**			
K			○							x					○							x				**x**		
Y												x												○			**x**	
Q		x					○												○									**x**

Substitution matrix for Scoring alignment:

A 20 x 20 matrix with a score for each possible pair of amino acids. Substitution matrices can be based on:

1) Chemical similarity of the amino acid side chains.
2) Genetic code scoring that depends on the minimum number of base changes needed to interconvert the codons for the two amino acids.
3) Observed substitutions in sequence of related proteins, e.g., BLOSUM62 for BLAST.

BLAST Blosum62 scoring matrix

The amino acids are represented along the left side and top using their one letter codes.

BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
Blocks Database = /data/blocks5.0/blocks.dat
Cluster Percentage: >= 62
Entropy = 0.6979, Expected = -0.5209

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	**9**	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	**5**	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	**5**	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	**6**	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	**8**	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	**4**	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	**5**	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	**5**	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	**5**	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	**11**	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	**7**	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	**4**

Insertions and Deletions:

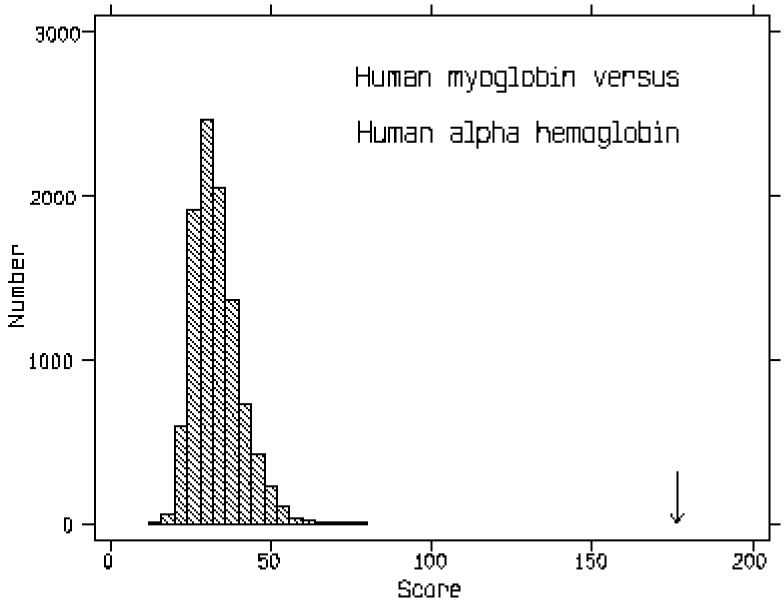
Gap opening penalty for creation of a gap and gap extension penalty for longer gaps give weighting terms that depend on the likelihood of an insertion or deletion and its length. Very similar proteins will have no gaps in their sequence alignments. Dissimilar proteins will have extensive and long gaps. The gap weights are empirical.

Significance of alignment:

A common and simple test to determine if the alignment of two sequences is statistically significant is a simple permutation test. This consists of

1. Randomly rearrange the order of one or both sequences
2. Align the permuted sequences
3. Record the score for this alignment
4. Repeat steps 1-3 a large number of times.

Doing this 10000 times gives a distribution of alignment scores that could be expected for random sequences with a similar amino acid content. If the actual alignment has a score much higher than that of the permuted sequences, then they must be homologous to some extent.



Plot of 10,000 alignment scores for the human myoglobin and alpha hemoglobin sequences.

The distribution is skewed - statistics based on a normal distribution would be strongly biased. The skew is expected since in each case the alignment algorithm is trying to maximize the score. The score for the alignment of the two actual sequences is 179 (indicated by the arrow). Obviously, myoglobin and haemoglobin are evolutionarily related and retain similar sequences. This alignment has a probability of less than 0.0001 of occurring by chance alone.