Sequence Alignment

Pratheeba Jeyananthan

Faculty of Engineering,
University of Jaffna

Lecture Plan

- Lecture Hours: 05 hours
- Lab Hours: 03 hours
- Reference: Understanding Bioinformatics by Marketa Zvelebil and Jeremy O. Baum

Introduction

- Invention of DNA sequencing technique in 1970s.
- Currently so many databases with nucleotide and protein sequences from a variety of organism.
- Those organisms include human, mouse, chimpanzee, fruit fly, yeast and numerous bacteria, archaea and viruses.
- Entries for nucleotide and protein sequences are in databases such as GenBank, dbEST and UniProt KB.
- Now the challenge is converting this sequence information into useful biological knowledge.
- To Find out more about newly determined sequence, we need to do the sequence analysis.

Key Stages in Sequence Analysis:

- One of the key stages in the most sequence analyses:
  - Alignment of different sequences to detect homology.
  - The comparison of a novel sequence with those in the databases to see whether there is any similarity between them.

- This chapter will cover,
  - Practical use of techniques and programs for general alignment, databases searching and pattern searching.
  - Alignment and analysis of protein sequences.

Why this?

- Identification of similar sequences has a multitude of applications.
  - For raw uncharacterized geniomic DNA sequences, comparison with sequences in a database can often tell you whether the sequence is likely to contain, or be part of a protein-coding gene.
  - Similarity search may retrieve a known gene or family of genes with a strong similarity to the new sequence.
  - This will provide the first clues to the type of protein the new gene encodes and its possible function.
  - Similarities in sequence can also help in making predictions about a protein's structure.
  - Sequences of proteins or DNAs from different organisms can also be compared in order to construct phylogenetic trees.
  - For the purpose of comparison, vast amount of DNA and protein sequences from various organisms are now available.
  - Although newly discovered sequences will share some or considerable similarity to sequence in the databases, there will still be many that are unique.

Principles of Sequence Alignment

- Devising ways of comparing sequences has never been straightforward, WHY?

Because,

- Vast amount of information now available for searching.
- Of the many ways DNA and protein sequences can change during evolution.
- Mutation and selection over millions of years can result in considerable divergence between present-day sequences derived from the same ancestral gene.
- Based on originally corresponding positions and the amino acids they encode can change as a result of point mutation, and the sequence lengths can be quite different as a result of insertions and deletions.
- Even more dramatic changes may also occur,
  - For example, Fusion of sequence from two different genes.
- Gene Duplications are common in eukaryotic genomes, and pseudogenes.

Pseudogenes

- What is a pseudogene?
  - In many cases mutation has disabled one copy of a gene so that it is either no longer expressed or if it is transcribed does not produce a functional protein.
- Pseudogenes are sequences in genomic DNA that have a similar sequence to known protein-coding genes but do not produce a functional protein.
- They are assumed to arise after gene duplication, when one of the gene copies undergoes mutation that either prevents its transcription or disrupts its protein-coding sequence.
- The human genome is estimated to contain up to 20,000 pseudogenes.
- As the pseudogene sequence is no longer under selection to retain protein function, it will generally accumulate further mutations at a higher rate than the functional gene.
- Despite this, many pseudogenes retain considerable sequence similarity to their active counterparts.

Sequence Alignment …

- On superficial inspection, such changes in gene sequence and length can effectively mask any underlying sequence similarity.
- To reveal it, the sequences have to be aligned with each other to maximize their similarities.
- This crucial step in sequence comparison is one of the main topic of this chapter.

Alignment …

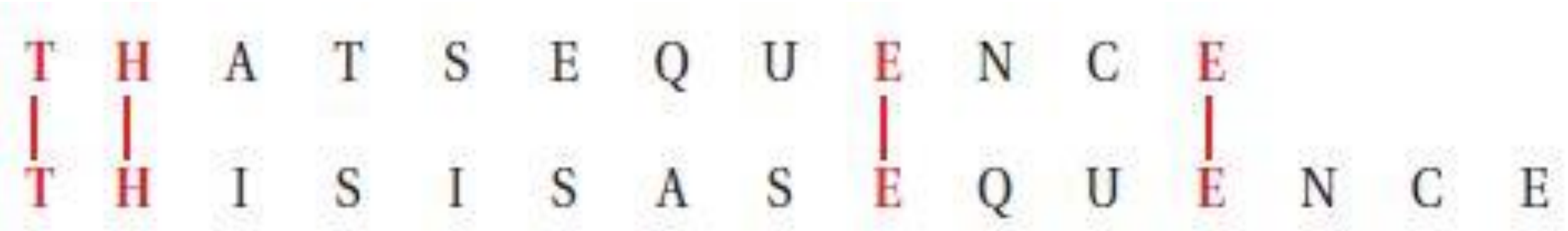
- As the result of mutation, even the sequences of the same protein or gene from two closely related species are rarely identical.
- Ideally, what we want to achieve when comparing sequences is to line them up in such a way that maximizing the similarity of aligned regions.
- To illustrate the general principle, take the two hypothetical amino acid sequences THISSEQUENCE and THATSEQUENCE.
- If we align them so that as many identical letters as possible pair up we get.

Alignment …

- Here, letters in red are identical.
- As we can easily see with such short and similar sequences, this alignment clearly identifies their strong similarity to each other.
- So far so good, but when sequences become more different from each other, they become more difficult to compare.
- How would we go about comparing the two sequences THATSEQUENCE and THISISASEQUENCE

Alignment …

- Here a mutation has led to the insertion of the three amino acids I, S, A into one of the original sequences.
- Simply lining them up from the beginning loses much of the similarity we can see exists.
- More subtly, because of the difference in length, it also creates false matches between noncorresponding positions.



Alignment …

- To get round this problem, gaps are introduced into one or both of the sequences so that maximum similarity is preserved.

- There is never just one possible alignment between any two sequences.
- The best one is not always obvious, especially when the sequences are not very similar to each other.
- At the heart of sequence-comparison and database-searching methods are,
  - algorithms for testing the fit of each alignment generated
  - giving it a quantitative score
  - and filtering out the unsatisfactory ones according to preset criteria.

Alignment and homology between sequences

- In all methods of sequence comparison, the fundamental question is whether the similarities perceived between two sequences are due to,
  - chance
  - thus of biological significance
  - due to the derivation of the sequences from a common ancestral sequence.
- Similarity versus homology
  - Similarity is simply a descriptive term telling you that the sequences in question show some degree of match.
  - Homology, in contrast, has distinct evolutionary and biological implications.
- Homologous genes are therefore genes,
  - derived from the same ancestral gene.
  - during their evolutionary history they will have diverged in sequence as a result of accumulating different mutations.

Homology

- Because homology implies a common ancestor, it can also imply a common function or structure for two homologous proteins.
  - can be a useful pointer to function if one of the proteins is known only from its sequence.

- Similar or identical aligned residues may simply be due to relatively recent divergence of the two sequences.
  - so care must be taken not to overestimate their functional importance.

- Moreover, mutation and selection can generate proteins with new functions but relatively little change in sequence.

- Therefore, sequence similarity does not always imply a common function.

Homology …

- An alignment of two sequences is, in effect, a hypothesis about which pairs of residues have evolved from the same ancestral residue.

- But an alignment in itself does not imply an evolutionary order of events.
  - two alternatives of homology and convergent evolution cannot usually be distinguished without additional information.

- Sequence comparison methods have to take account of such factors as,
  - the types of mutation that occur over evolutionary time
  - differences in the physicochemical properties of amino acids and their role in determining protein structure and function
  - the selective pressures that result in some mutations being accepted and others being eliminated.

- One has to consider the evolutionary processes that are responsible for sequence divergence and find a way to include the salient features in practicable schemes for testing the goodness of fit of the alignment

Homology …

- These must be quantitative and hence involve a score.
- Such scoring schemes can then be incorporated in algorithms designed to generate the best possible alignments.
- Finally, ways must be found to discriminate between fortuitously good alignments and those due to a real evolutionary relationship.
- For most purposes, comparisons of protein sequences show up homology more easily than comparisons of the corresponding DNA sequences.
- There are many reasons for this greater sensitivity.
- First, there are only four letters in the DNA alphabet compared to the 20 letters in the protein alphabet, and so a DNA sequence, of necessity, provides less information at each sequence position than does a protein sequence.

Detecting Homology

- In other words, there is a much greater probability that a match at any one position between two DNA sequences will have occurred by chance.

- Therefore, the degree of similarity, as judged by some appropriate quantitative score, needs to be greater between DNA sequences than between protein sequences for the alignment to be of importance.

Substitution matrices and scoring

- The large numbers of sequences that can be examined for similarity nowadays oblige us to use automated computational methods to judge the quality of an alignment, at least as an initial filter.

- It is possible for two sequences to be aligned in a variety of different ways, including the insertion of gaps to improve the number of matched positions.

- How does one objectively determine which is the best possible alignment for any given pair of sequences?

- In practice, this is done by calculating a numerical value or score for the overall similarity of each possible alignment so that the alignments can be ranked in some order.

Scoring

- We can then work on the basis that alignments of related sequences will give good scores compared with alignments of randomly chosen sequences.

- The correct alignment of two related sequences will ideally be the one that gives the best score.

- The alignment giving the best score is referred to as **the optimal alignment**, while others with only slightly worse scores are often called **suboptimal alignments.**

- No one has yet devised a scoring scheme that perfectly models the evolutionary process, which is so complex that it defies any practical method of modeling.

- The implication of this is that the best-scoring alignment will not necessarily be the correct one, and conversely, that the correct alignment will not necessarily have the best score.

Scoring …

- However, the scoring schemes now in common use, are generally reliable and useful in most circumstances, as long as the results are treated with due caution and regard for biological plausibility.

- In principle, a scoring scheme can either measure similarity or difference.

- The simplest way of quantifying similarity between two sequences is percentage identity.

- Identity describes the degree to which two or more sequences are actually identical at each position.

- It simply measured by counting the number of identical bases or amino acids matched between the aligned sequences.

Percentage Identity

- **Percentage** or **percent identity** is obtained by dividing the number of identical matches by the total length of the aligned region and multiplying by 100.
- For the THATSEQUENCE/THISISASEQUENCE comparison, for example, has a percentage identity score of 68.75%, 11 matches over a total length of 16 positions, including the gaps.
- One might think that an alignment of completely unrelated sequences would have a percentage identity of zero.
- However, as there are only four different nucleotides in nucleic acid sequences, and only 20 different amino acids in protein sequences there is always a small but finite probability for any aligned sequences that identical residues will be matched at some positions.
- Because there are often hundreds of residues in a protein sequence and thousands in a nucleotide sequence, unrelated sequences are expected to align matches at several positions.
- The length of the sequence matters: a 30% identity over a long alignment is less likely to have arisen by chance than a 30% identity over a very short alignment.
- Statistically rigorous methods have been devised to measure the significance of an alignment.

Dot-matrix or dot-plot

- A dot matrix or **dot-plot** is one of the simplest ways to compare sequence similarity graphically, and can be used for both nucleotide and protein sequences.

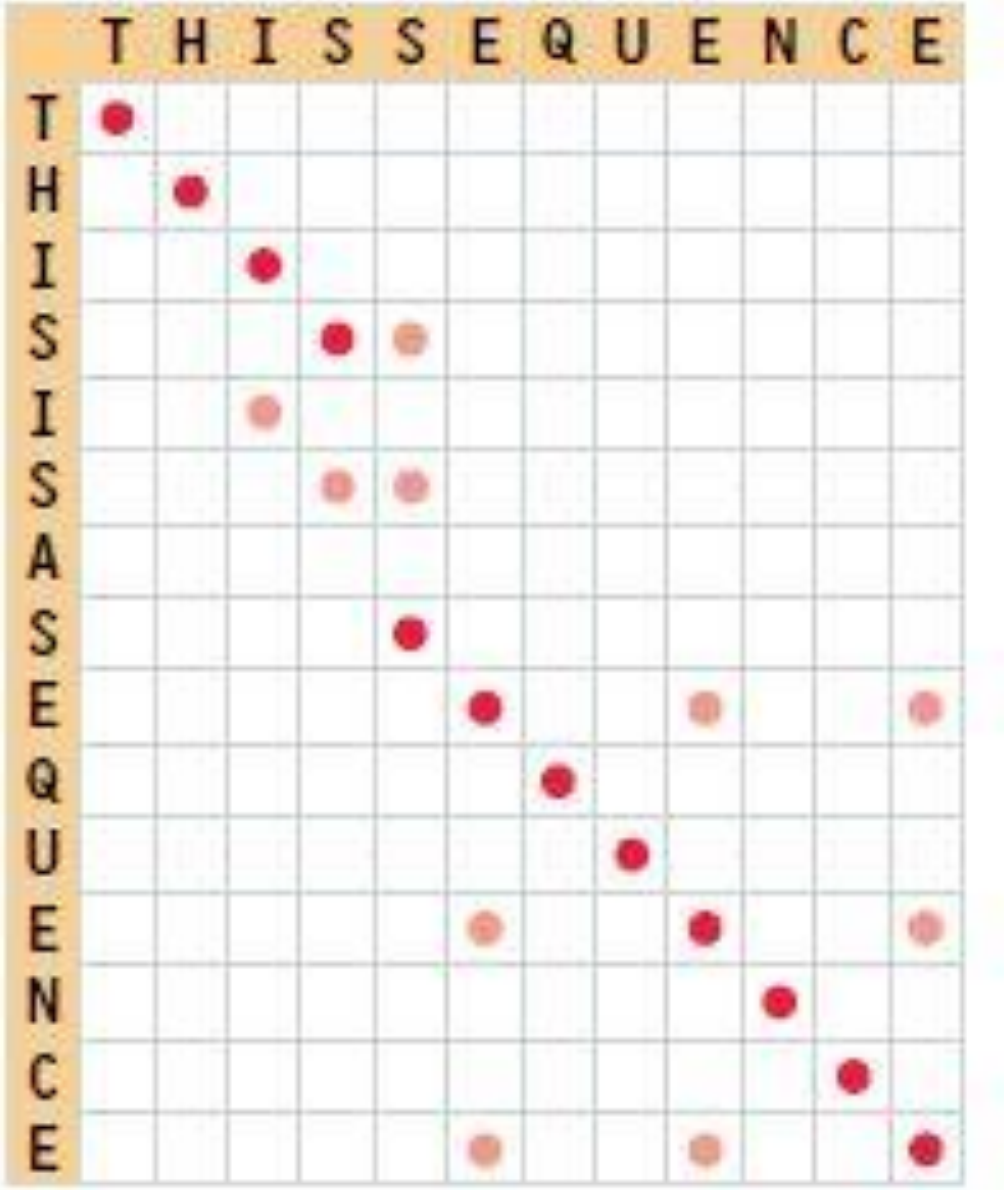
- To compare two sequences X and Y, one sequence is written out vertically, with each residue in the sequence represented by a row, while the other is written horizontally, with each residue represented by a column.

- Each residue of X is compared to each residue of Y (row to column comparison) and a dot is placed where the residues are identical.

- In the simplest scoring system, identical residues are scored as 1 and nonidentical residues as 0, and dots are placed at all positions that contain a 1.

- For example, if we take the pair THISSEQUENCE/THISISASEQUENCE pair, then a simple dot-plot will look like as follows:

Dot-Plot Representation



Dot-plot …

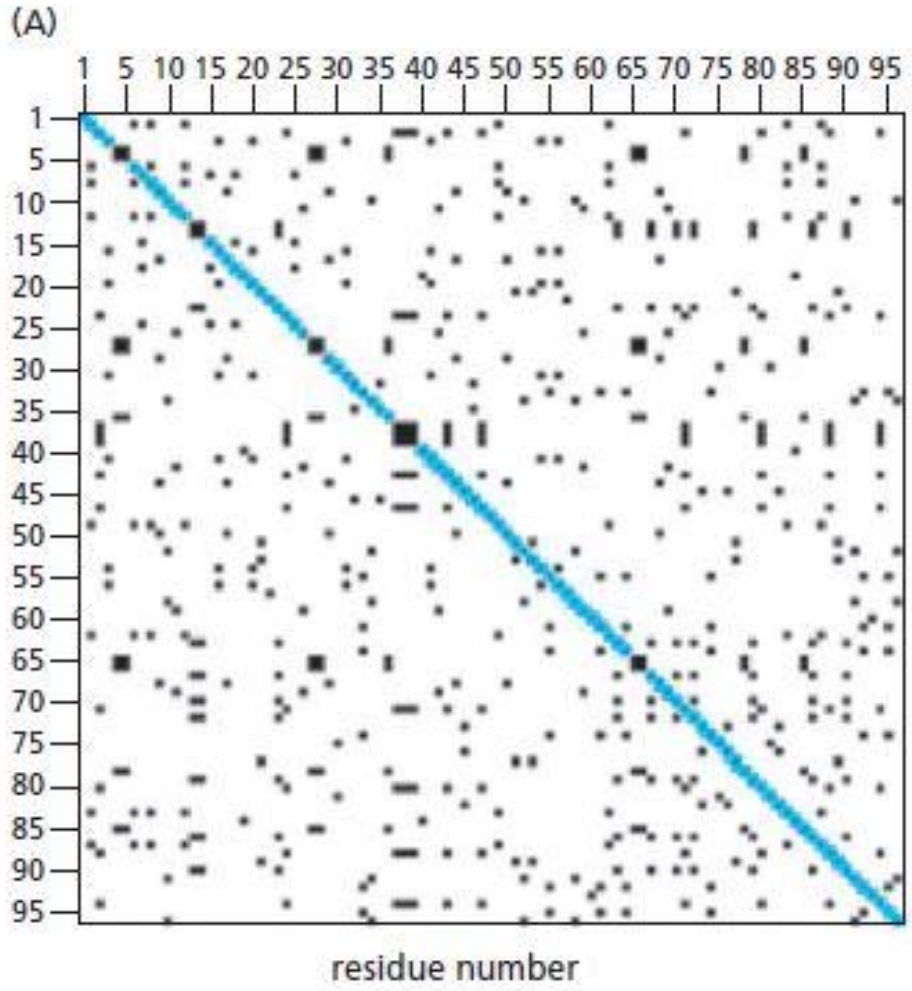
- The dots in red, which form diagonal lines, represent runs of matched residues.
- The pink dots scattered either side of the diagonals are the same residues found elsewhere in the sequence.
- The diagonals are interrupted by a few cells, where a gap has been inserted.
- Dot-plots can be useful for identifying intrasequence repeats in either proteins or nucleic acids.
- However, dot-plots suffer from background noise.
- To distinguish dot-patterns arising from background noise from significant dot-patterns it is usually necessary to apply a filter.
- The most widely used filtering procedure uses overlapping fixed-length windows and requires that the comparison achieve some minimum identity score summed over that window before being considered.

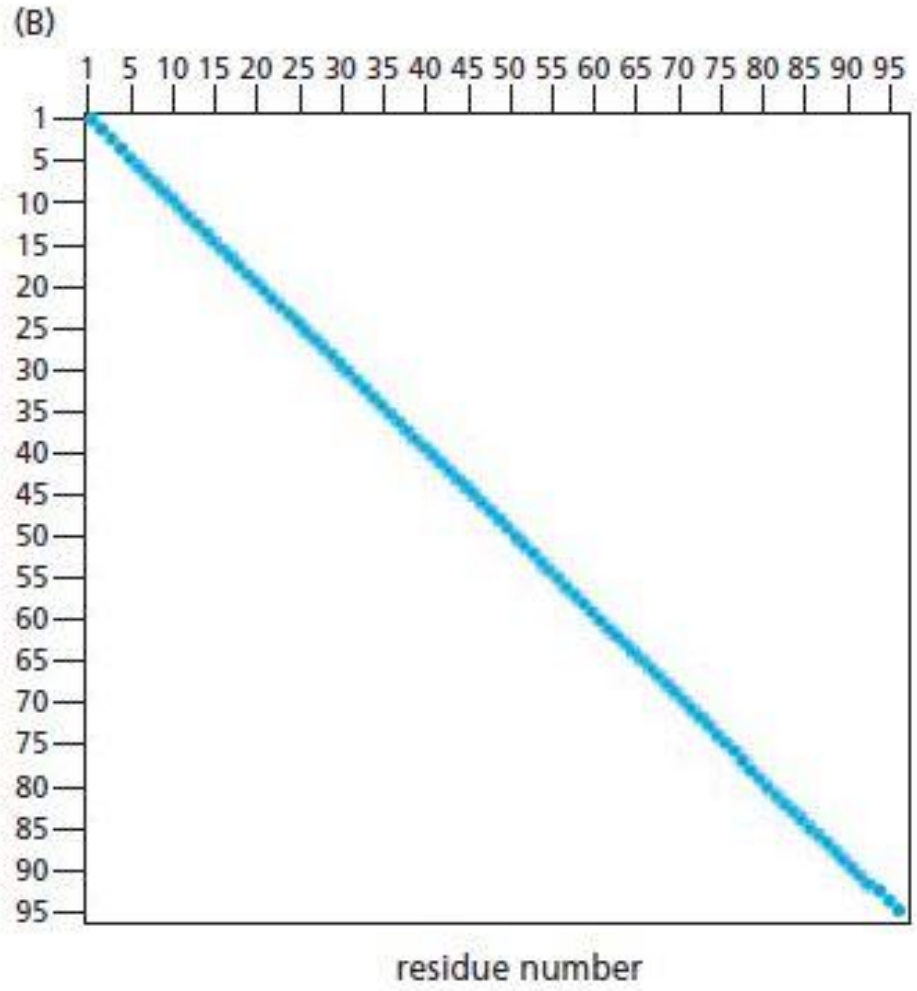
Dot-plot between two SH_2 sequences

- The SH_2 or Src-homology 2 domain is a small domain of about 100 residues found in many proteins involved in intracellular signaling in mammalian cells.
- It gets its name from the protein tyrosine kinase Src, where it was first found.
- Following figure has a window length of 1; in other words, every residue is considered individually.
- Diagonal line indicating matched identical residues is clear and unbroken.
- one would expect from a comparison of two identical sequences, there is still a certain amount of background noise detracting from the result, as most types of amino acid occur more than once in the sequence.
- Fig B shows the same comparison with a window of 10 residues and a minimum score for each window set to 3.
- Only the main diagonal is now seen, representing the one-to-one matching of the identical sequences

Dot-plot SH_2 protein

Two views of dot-plot representations of an SH2 sequence compared with itself. (A) Unfiltered dot-plot (window length = 1 residue). The identity between the two sequences is shown by the unbroken identity diagonal. Nevertheless, there is still background noise. (B) Dot-plot of the same sequence comparison with a window of 10 residues and a minimum identity score within each window set to 3. The background noise has all been removed, leaving only the identity diagonal.





Dot-plot Window size

- Most dot-plot software provides a default window length and this is sufficient for an initial analysis.
- But one can use the window length to greater effect by varying it depending on what one is searching for.
- Window length can be set, for example, to the length of an exon when comparing coding sequences, or to the size of an average secondary structure within a protein when looking for structural motifs.
- When searching for internal repeats, the length of the repeat can be used to cut out background noise.
- In addition, rather than using 0 and 1 as the scores for nonidentical and identical residues, other values can be used and the score can be varied depending on the type of residues involved.

Genuine matches do not have to be identical

- Although it is the simplest alignment score to obtain, and can be very useful as a quick test of the quality of an alignment, percentage identity is a relatively crude measure and does not give a complete picture of the degree of similarity of two sequences to each other, especially in regard to protein sequences.

- For example, simply scoring identical matches as 1 and mismatches as 0 ignores the fact that the type of amino acid involved is highly significant.

- In particular, certain nonidentical amino acids are very likely to be present in the same functional position in two related sequences, and thus are likely to represent genuine matches.

- This is chiefly because certain amino acids resemble each other closely in their physical and/or chemical properties and can thus substitute functionally for each other.

- Mutational changes that replace one amino acid with another having similar physicochemical properties are therefore more likely to have been accepted during evolution.

Percent Similarity

- So pairs of amino acids with similar properties will often represent genuine matches rather than matches occurring randomly.

- The simplest way of taking this into account is simply to count such similar pairs of amino acids as matches, and to refer to the score as **percent similarity.**

- In the familiar example sequences below, red is used to indicate residues that are similar but not identical.

- Here the sequences have been realigned to take into account similarity as well as identity.

```
T H I S I S A S E Q U E N C E
| |           | | | | | | | |
T H A T - - - S E Q U E N C E
```

Percent Similarity …

- Isoleucine (I) and alanine (A) are similar as they are both hydrophobic, whereas serine (S) and threonine (T) both have an -OH group in their side chain and are polar.

- Not all similar amino acid pairs are equally likely to occur, however, and more sophisticated measures of assessing similarity are more commonly used.

- In these, each aligned pair of amino acids is given a numerical score based on the probability of the relevant change occurring during evolution.

- In such scoring schemes,
  - pairs of identical amino acids are assigned the highest score;
  - then, pairs of amino acids with similar properties (such as isoleucine and leucine) score more highly than those with quite different properties (such as isoleucine and lysine) which are rarely found in corresponding positions in known homologous protein sequences.

Minimum percentage identity

- What is the minimum percentage identity that can reasonably be accepted as significant?
- Burkhard Rost analyzed more than a million alignments of pairs of protein sequences for which structural information was available to find a cut-off for the level of sequence identity below which alignment becomes unreliable as a measure of homology.
- He found that 90% of sequence pairs with identity at or greater than 30% over their whole length were pairs of structurally similar proteins.
- Given both sequence and structural similarity, one can usually be confident that two sequences are homologous, so 30% sequence identity is generally taken as the threshold for an initial presumption of homology.
- Below about 25% sequence identity, however, Rost found that only 10% of the aligned pairs represented structural similarity.
- The region between 30% and 20% sequence identity has been called the twilight zone, where homology may exist but cannot be reliably assumed in the absence of other evidence.
- Even lower sequence identity (<20%) is referred to as the midnight zone.

Different ways of scoring an alignment

- The function of an alignment score is to provide a single numerical value for the degree of similarity or difference between two sequences.
- Most current applications measure similarity, and in this case the highest scores are best.
- A few applications, particularly those used for generating phylogenetic trees, use a score related to sequence difference, usually known as a distance, in which case the most closely related sequences give alignments with the lowest scores.
- The measure of difference between two homologous sequences from different species is sometimes called the genetic or evolutionary distance.
- There is no a priori reason why residue pair alignment scores cannot be negative, for example to represent especially unlikely alignment.

Different ways of scoring an alignment …

- In fact, some of the popular techniques require scores that can be negative, and most commonly used schemes have both positive and negative scores for pairs of residues.
- Scoring schemes have to represent two salient features of an alignment.
- On the one hand, they must reflect the degree of similarity of each pair of residues; that is, the likelihood that both are derived from the same residue in the presumed common ancestral sequence.
- On the other hand, they must assess the validity of inserted gaps.

Substitution Matrices

- Substitution matrices are used to assign individual scores to aligned sequence positions.
- For alignments of protein sequences, the score assigned to each aligned pair of amino acids is generally determined by reference to a substitution matrix, which defines values for all possible pairs of residues.
- Various types of substitution matrices have been used over the years.
- Some were based on theoretical considerations such as the number of mutations that are needed to convert one amino acid into another, or similarities in physicochemical properties.
- The choice of which substitution matrix to use is not trivial because there is no one correct scoring scheme for all circumstances.

Substitution Matrices …

- The most important scoring matrices is described below with general guidance as to which one to use when.
- When an alignment is made, each aligned amino acid pair is given a score from the substitution matrix.
- These scores are then summed to give the overall score (S) of the alignment.
- For example (sing the BLOSUM-62 matrix):

Seq1:	T	H	I	S	S	E	Q	U	E	N	C	E
Seq2:	T	H	A	T	S	E	Q	U	E	N	C	E
Score:	5	8	–1	1	4	5	5	0	5	6	9	5

- Therefore the overall score S for this alignment equals 52. The BLOSUM matrices are described in more detail below.

The BLOSUM-62 matrix

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

The PAM120 substitution matrix

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

PAM (Point Accepted Mutations)**substitution matrix**

- A commonly used set of substitution matrices is based on the observed amino acid substitution frequencies in alignments of homologous protein sequences.

- Have been found to be superior to substitution schemes that use only the physicochemical similarities of amino acids, as they use real data to model the evolutionary process.

- The sequences used to generate these matrices were all very similar, allowing the alignment to be made with confidence.

- In addition, the high similarity meant that there was a high probability that amino acid differences at an alignment position were due to just a single mutation event, over a short period of time, since it is unlikely that more than one mutation would occur at the same site.

- There is more than one such matrix and each matrix corresponds to a particular quantity of accepted mutations — mutations that have been retained in the sequence.

BLOSUM substitution matrix

- The BLOSUM matrix is another very commonly used amino acid substitution matrix that depends on data from actual substitutions.
- First, a large set of aligned highly conserved short regions was generated from analysis of the protein-sequence database SWISS-PROT.
- The sequences were then clustered into groups according to similarity, so that sequences were grouped together if they exceeded a specified threshold for percentage identity.
- Substitution frequencies for all possible pairs of amino acids were then calculated between the clustered groups (without the construction of phylogenetic trees) and used to compute BLOSUM (BLOck SUbstitution Matrix) scores.
- Various BLOSUM matrices are obtained by varying the percentage cut-off for clustering into similarity groups.
- For example, the commonly used BLOSUM-62 matrix was derived using a threshold of 62% identity.

Choice of substitution matrix

- Some scoring matrices have been designed to work well in special situations.
- For example, the matrices SLIM (ScoreMatrix Leading to Intra-Membrane) and PHAT (Predicted Hydrophobic And Transmembrane matrix) are especially designed for membrane proteins.
- In 2006, there were 94 matrices collected in a database list called AAINDEX and searchable at GenomeNet.
- As well as the degree of evolutionary distance, the length of the sequences to be aligned must be taken into account when choosing a suitable matrix.
- Short sequences need to use matrices designed for short evolutionary time scales, such as PAM40 or BLOSUM-80.
- Longer sequences of 100 residues or more can use matrices intended for use with longer evolutionary time scales (such as PAM250 and BLOSUM-50.

Inserting Gaps

- Gaps inserted in a sequence to maximize similarity with another require a scoring penalty.
- Homologous sequences are often of different lengths as the result of insertions and deletions (indels) that have occurred in the sequences as they diverged from the ancestral sequence.
- Their alignment is generally dealt with by inserting gaps in the sequences to achieve as correct a match as possible.
- To signify that an insertion or deletion has occurred, a letter or stretch of letters in one sequence is paired up with blank spaces (usually indicated by hyphens) inserted into the other sequence to achieve a better match.
- Gaps must be introduced judiciously: forcing two sequences to match up simply by inserting large numbers of gaps will not reflect reality and will produce a meaning-less alignment.
- To place limits on the introduction of gaps, alignment programs use a gap penalty: each time a gap is introduced, the penalty is subtracted from the score, decreasing the overall score of the alignment.

Inserting Gaps …

- If a low gap penalty is chosen, then more and larger gaps will be inserted.
- Therefore, if you are searching for sequences that are a strict match for your query sequence, the gap penalty should be set high.
- This will often retrieve a region, or regions, of very closely related sequence.
- If you are searching for similarity between distantly related sequences, the gap penalty should be set low.
- Note that suitable gap-penalty values may be different with different substitution matrices.
- It is advisable to start, when possible, with a combination of matrix and gap penalties that have been reported to give optimal performance.

Pairwise alignments of the PI3-kinase p110a and a cAMP-dependent protein kinase

Inserting Gaps …

- In the second alignment (Figure B) the gap penalty was set very low; the effect is that many more gaps are inserted and the number of matched amino acids is increased (identities are shown in green).

- Although there are more matched residues in the alignment with low gap penalties, this does not necessarily mean that it is more accurate.

Dynamic Programming Algorithms

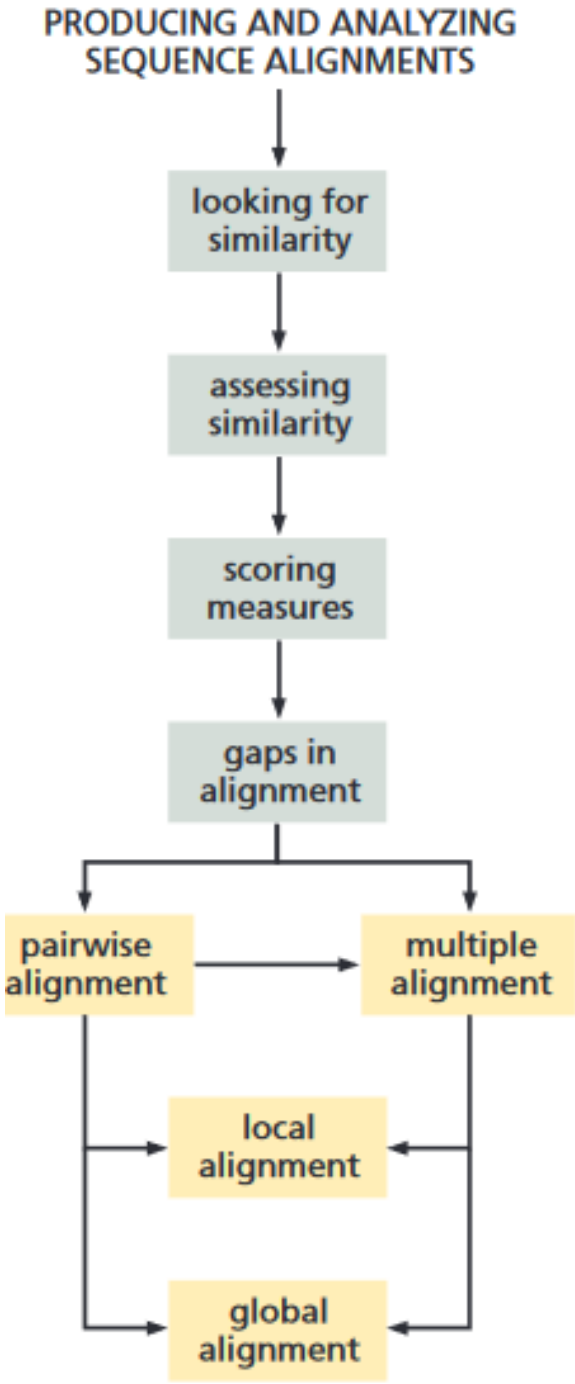
- Dynamic programming algorithms can determine the optimal introduction of gaps.
- In practice, it is nearly always necessary to insert gaps into sequences when aligning them.
- The most obvious way of finding the best alignment with gaps would be to generate all possible gapped alignments, find the score for each, and select the highest-scoring alignment.
- This would be enormously time consuming, however.
- It only became practicable to incorporate gaps into an alignment with the development of dynamic programming algorithms.
- The name “dynamic programming” reflects the fact that the precise behavior of the algorithm is established only when it runs (in other words, dynamically) because it depends on the sequences being aligned.

Dynamic Programming Algorithms …

- The first algorithm to use dynamic programming for sequence comparison was that of S. B. Needleman and C. D. Wunsch, published in 1970.
- Their technique is still the core of many present-day alignment and sequence-searching methods.
- In their method, gaps, regardless of length, have an associated penalty score; newer methods use more complicated gap penalties.
- The actual values of the gap scores can be varied depending on the type of scoring matrix being used.
- One rule always followed is that gaps can never be aligned with each other.
- The basic concept of a Needleman–Wunsch-type algorithm is that comparisons are made on the basis of all possible pairs of amino acids that could be made between the two sequences.
- All possible pairs are represented as a two-dimensional matrix, in which one of the sequences to be aligned runs down the vertical axis and the other along the horizontal axis.
- All possible comparisons between any number of pairs are given by pathways through the array, each of which can be scored.

Types of Alignment

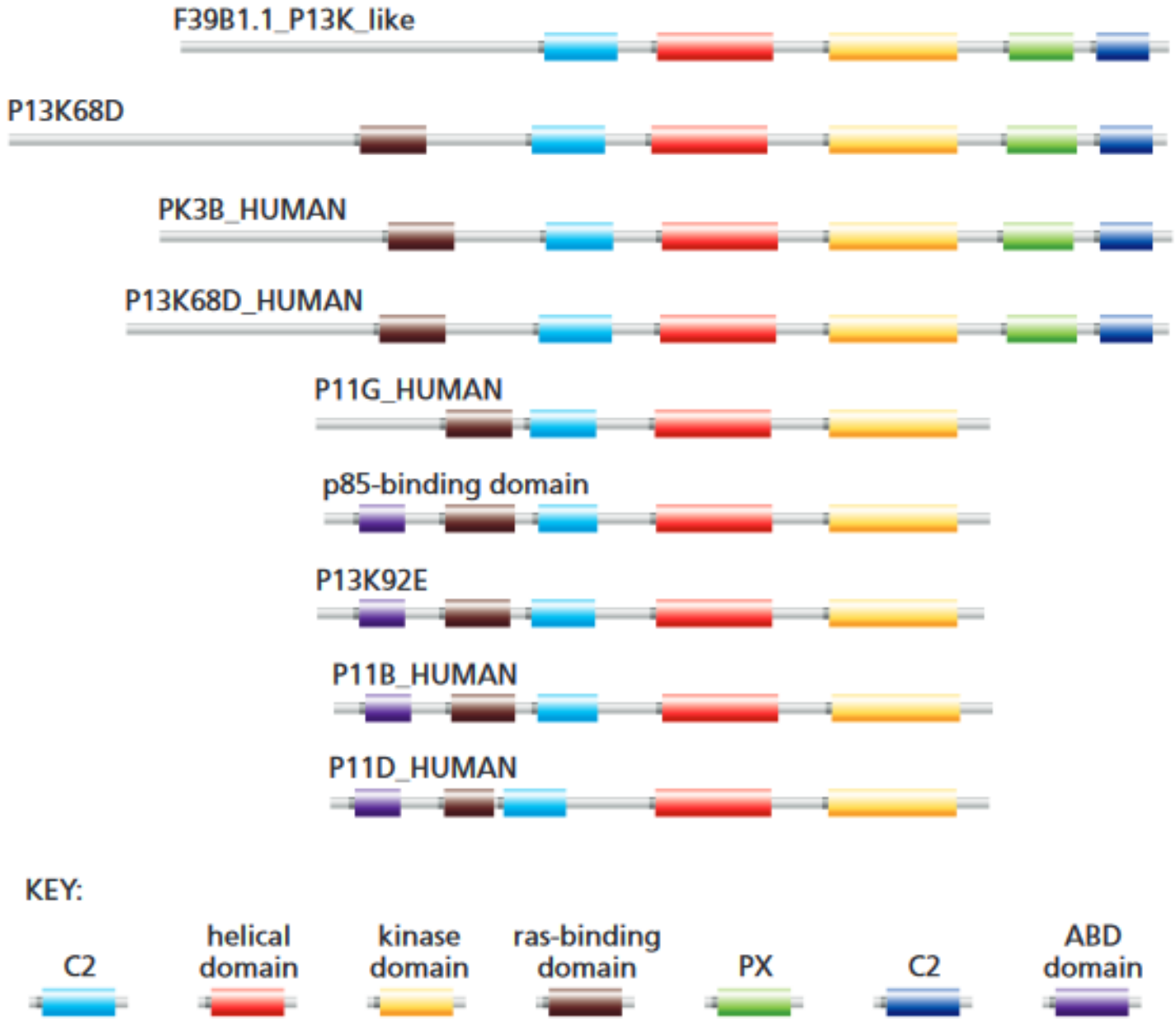
- Different kinds of alignments are useful in different circumstances.
- The general principles outlined in the previous sections can be used to make different types of alignment, see the following diagram:
- Two closely related homologous sequences will generally be of approximately the same length, so that their alignment will cover the full range of each sequence.
- This is referred to as a global alignment, and is generally the appropriate one to use when you want to compare or find closely related sequences that are similar over their whole length.



Types of Alignment …

- On the other hand, there are many cases where only parts of sequences are related.
- A simple example is the amino acid sequences of two proteins each consisting of two domains, with only one domain common to both proteins and the other domains completely unrelated.
- In this case, the only meaningful alignment will be a local alignment of the shared domain.
- Looking only at global alignments may not reveal the limited but important similarity between the sequences.
- This is particularly the case for comparisons between multidomain proteins, such as PI3-kinases, which consist of a number of small protein domains strung together (Below figure).
- Local alignment programs are therefore useful for detecting shared domains in such proteins.
- When searching through a sequence database with a query sequence from an unknown protein, local alignment is a very useful tool to use initially.

PI3-kinase is a multidomain protein



Types of Alignment …

- Once sequences with regions of high similarity are found using local alignment, global alignment can be used to align the rest of the sequence that is not so similar.
- Local alignment is also a good tool for identifying particular functional sites from which sequence patterns and motifs can be derived.
- A widely used local alignment algorithm is the Smith–Waterman algorithm, which is a modification of the Needleman–Wunsch algorithm.
- Instead of looking at each sequence in its entirety, which is what the Needleman–Wunsch algorithm does, the Smith–Waterman method compares segments of all possible lengths and chooses the segment that optimizes the similarity measure.
- The scoring matrix used must include both positive and negative scores, and only alignments with a positive total score are considered.
- Therefore, if on extending the alignment at a particular step none of the possible alignments has a positive score, all previous alignments are rejected, and new ones are considered starting from that point.
- This makes the calculation sensitive to the precise match and mismatch scores and gap penalties.

Local versus Global alignment

(A) local

```
PI3-kinase  DRHNSNIMVKDDGQLFHIDFG
cAMP PK     DLKPENLLIDQQGYIQVTDFG
```

(B) global

```
                    10                  20                  30                  40                  50
PI3-kinase  HQLGNLR--LEECRI---MSSAKRPLWLNWENPDIMSELLFQNNEIIFKNGDDLRQDMLT
cAMP PK     GNAAAAKKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVML-
                   10                  20                  30                  40                  50

                60                  70                  80                  90                  100                 110
PI3-kinase  LQIIRIME--NIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQ-IQCKGGLKGAL
cAMP PK     ---VKHMETGNHYAMKILDKQKVVK--------LKQIEHTLNEKRILQAVNFPFLVKLEF
               60                  70                  80                  90                  100

                        120                 130                 140                 150                 160
PI3-kinase  QFNSHT-LHQWLKDKNKGEIYDAA--IDLFTRSCAGYCVATFILGIGDRHNSNIMVKD-D
cAMP PK     SFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLK
           110                 120                 130                 140                 150                 160

                170                 180                 190                 200                 210                 220
PI3-kinase  GQLFHIDFGHFLDHKKKKFGYKRERVP-----FVLTQDFL---IVISKGAQECTKTREFE
cAMP PK     PENLLIDQQGYI--QVTDFGFAK-RVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALG
           170                 180                 190                 200                 210                 220

                        230                 240                 250                 260                 270
PI3-kinase  RF-QEMC--YKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEA
cAMP PK     VLIYEMAAGYPPFFA-DQPIQIYEKIVSGKVR--FPSHFSSDLKDLLRNLLQVDLTKR--
               230                 240                 250                 260                 270                 280

            280                 290                 300                                             310
PI3-kinase  LEYFMKQMNDAHHGGWTTKMDWI----------------------FHTIKQHALN----
cAMP PK     FGNLKNGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXIN
                       290                 300                 310                 320                 330                 340
```

Types of Alignment …

- Local versus global alignment of the complete protein sequences of the bovine PI3-kinase p110a and the cAMP-dependent protein kinase shown in above figure using the Web-based programs ALIGN (global) and LALIGN (local).

- Although these proteins share structural homology within the core kinase catalytic domain, there is very little sequence homology.

- Figure A shows that local alignment of the catalytic domains has identified one important conserved region, out of five regions that were aligned.

- Figure B shows that, in this case, a global alignment fails to identify this region.

- For both global and local alignments, methods exist for making pairwise alignments, that is, the alignment of just two sequences, and for making multiple alignments, in which more than two sequences are aligned with each other.

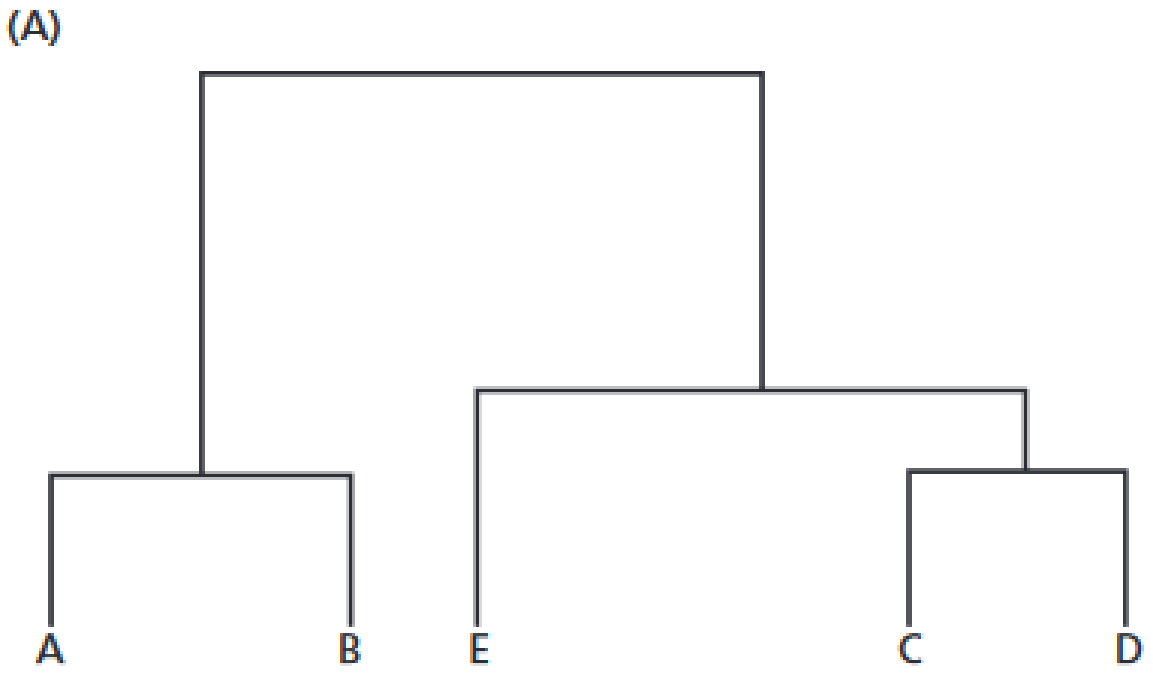
Types of Alignment …

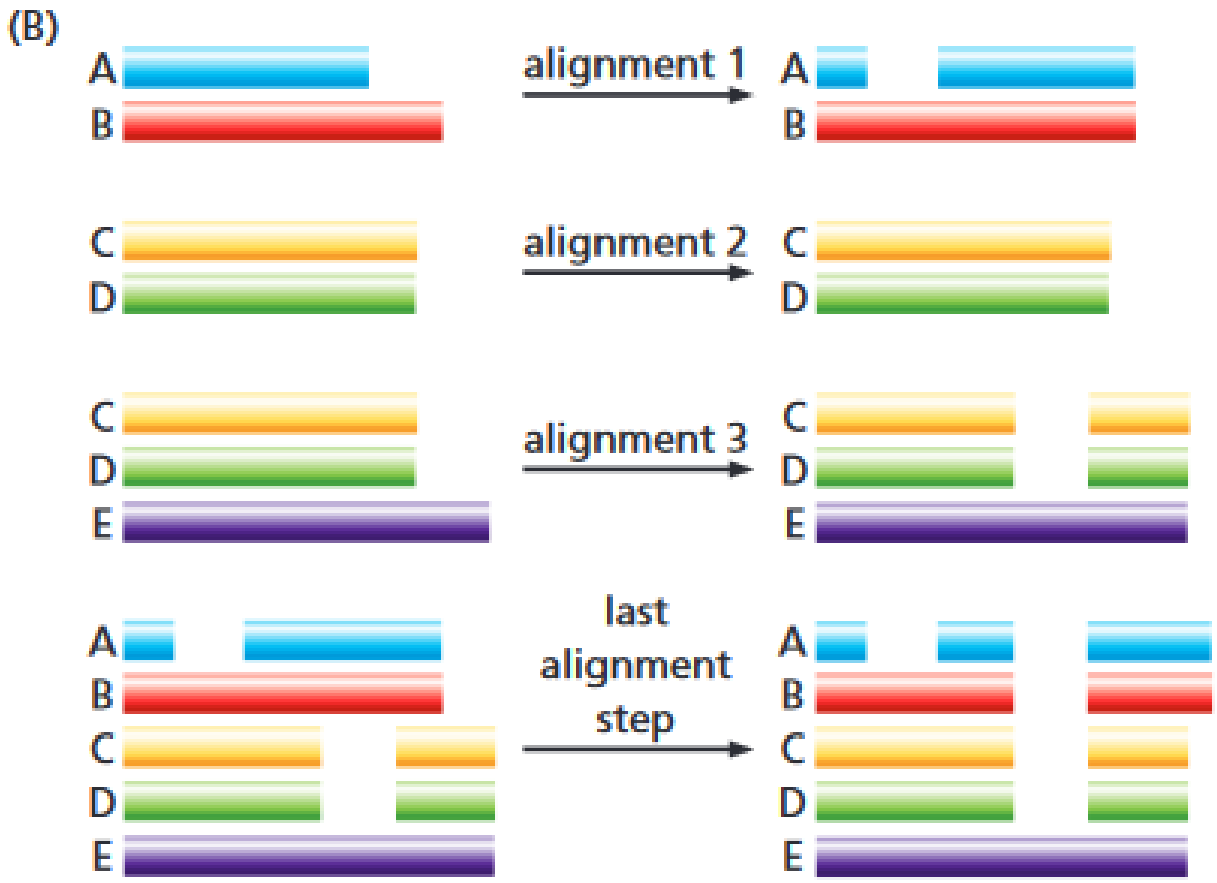
- In this part of the chapter, we have mainly used examples of pairwise alignments to illustrate the general principles of alignment scoring and quality assessment.
- Multiple alignment introduces yet another dimension to the computational problems of alignment.
- Multiple alignments can be used to find:
  - interesting patterns characteristic of specific protein families
  - to build phylogenetic trees
  - to detect homology between new sequences and existing families
  - to help predict the secondary and tertiary structures of new sequences
- In general, the alignment of multiple sequences will give a more reliable assessment of similarity than a pairwise alignment.
- The reason for this is that ambiguities in a pairwise comparison can often be resolved when further sequences are compared.
- Multiple alignment provides more information than pairwise alignment on the individual amino acid positions, such as the overall similarity and evolutionary relationships.

Multiple alignments can be constructed by several different techniques

- A number of methods are available for generating multiple alignments.
- One of these is an extension of the dynamic programming method, so that instead of a two-dimensional matrix for a pair of sequences, an alignment of n protein sequences uses an n-dimensional matrix.
- However, this is limited by the prohibitively large computational requirement of the algorithm, and none of the examples discussed below uses this technique.
- Other methods, while often using dynamic programming to align pairs of sequences, use other techniques to combine these together into one multiple alignment.
- Tree or hierarchical methods of multiple alignment are widely used, for example in the multiple alignment program ClustalW.
- This method first compares all the sequences in a pairwise fashion, then performs a cluster analysis on the pairwise data to generate a hierarchy of sequences in order of their similarity (Figure A).
- The hierarchy is often referred to as the guide tree.

Tree Method for the multiple alignment of sequences A, B, C, D and E





Multiple alignments can be constructed by several different techniques …

- A multiple alignment is then built based on the guide tree by first aligning the most similar pairs, then aligning the other sequences with these pairs until all the sequences have been aligned (Figure B).

- However there are problems with it in that any errors in the initial alignments cannot be corrected later as new information from other sequences is added.

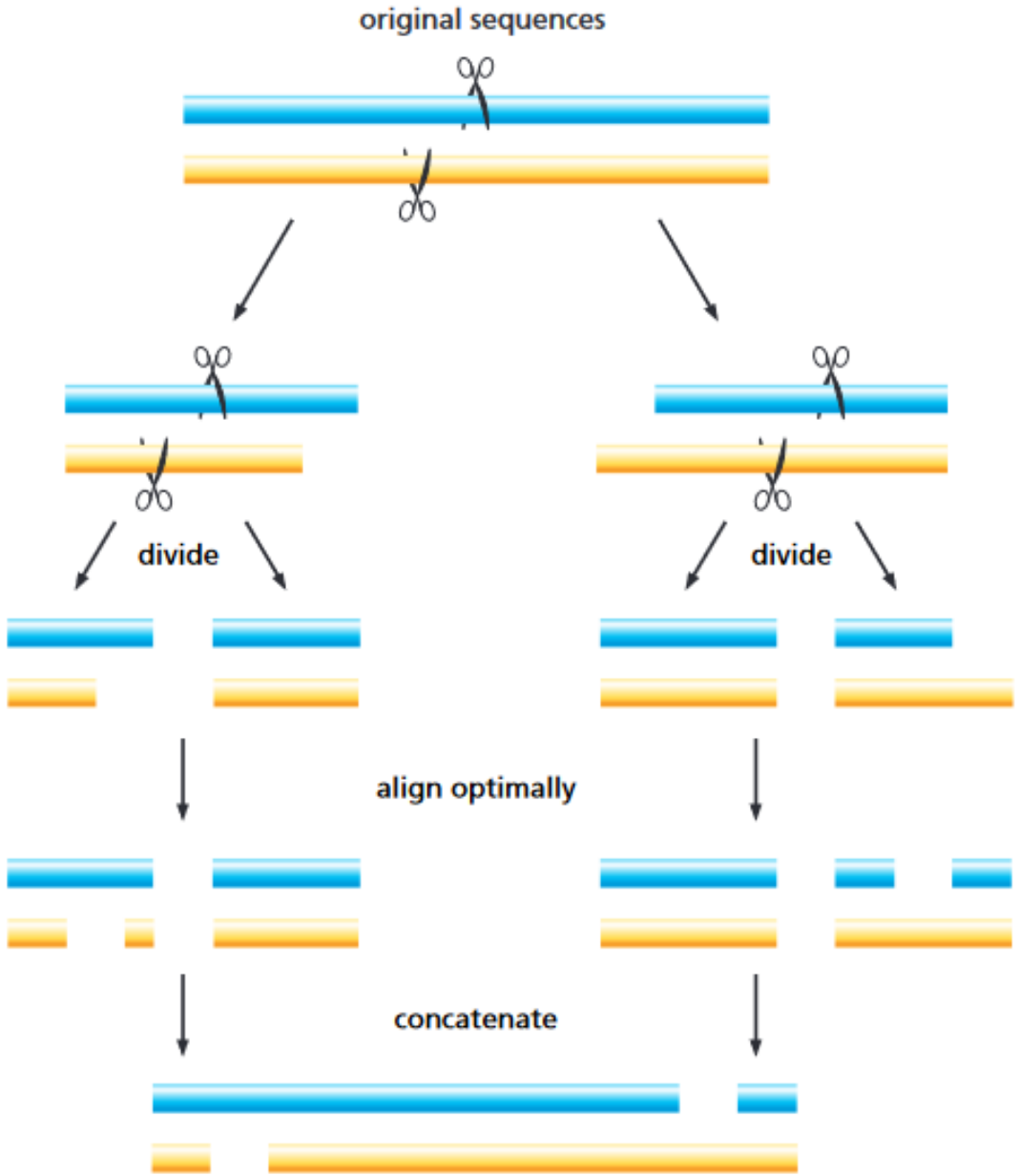
- This difficulty has been avoided in iterative or stochastic sampling procedures as in the Barton and Sternberg program.

- Other methods for building multiple alignments include the segment method, the consensus method, and the divide-and-conquer method.

- In the divide-and-conquer alignment, the sequences are first cut several times to reduce the length of the sequences to be aligned, the cut sequences are then aligned, and they are finally concatenated into a multiple alignment (Figure).

- Initially, each sequence is divided into two segments at a suitable cut-position somewhere close to the midpoint of the sequences.

The divide-and-conquer method of multiple alignment



Multiple alignments can be made by combining a series of local alignments

- DIALIGN is a relatively recent method for multiple alignment developed by Burkhard Morgenstern and colleagues.
- Whereas standard alignment programs such as ClustalW compare residues one pair at a time and impose gap penalties, DIALIGN constructs pairwise and multiple alignments by comparing whole ungapped segments several residues long.
- The alignment is then constructed from pairs of equal-length gap-free segments, which are termed diagonals because they would show up as diagonal lines in the respective pairwise comparison matrices.
- The segment length varies between diagonals.
- Many diagonals overlap, and the program has to find a set that can be combined into one consistent alignment.
- As the segments are gap-free there is no need to use a gap-penalty parameter.
- Every diagonal is given a weight reflecting the degree of similarity between the two segments involved.
- The overall score of an alignment is the sum of the weights of all the diagonals, and the program finds the alignment with the maximum score.

Multiple alignments can be made by combining a series of local lignments

- A threshold can be set so that diagonals are considered only if their weights exceed this threshold, so that regions of lower similarity are ignored.

- As DIALIGN is a local alignment method it may not align the whole sequence, and may align several blocks of residues with unaligned regions between them.

- Below figure illustrates the alignment of five SH2 domain sequences using ClustalW, DIALIGN, and the divide-and-conquer algorithm (DCA) methods compared with the structural/functional alignment from BAliBase, which can be considered accurate.

- All three methods fail to some extent to align the residues of the first helix correctly, inserting a gap.

- ClustalW does slightly worse in this region by splitting the helix, but is better in conserving the integrity of the second core block around the FLVR region important for binding.

- DCA does not align the last helix as well as ClustalW or DIALIGN.

- However, all the alignment programs are generally good and useful in that they often produce alignments very close to the correct ones based on extra information, such as those found in BALiBase.

Known structural alignments can be useful in checking sequence alignments

(A) structural/functional alignment from BAliBase

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRARD--NNGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGK-KFDTLWQLVEHYSYKA------DGLLRVL-TVPCQK
1gri  EMKPHPWFFGKIPRAKAEEML-SKQRHDGAFLIRESES-APGDFSLSVKFGNDVQHFKVLRDGAGKYFL-WVV-KFNSLNELVDYHRSTS-VSRNQQIFLRDIEQVPQQ-
1aya  ---MRRWFHPNITGVEAENLLLTRG-VDGSFLARPSKS-NPGDFTLSVRRNGAVTHIKIQN--TGDYYDLYGGEKFATLAELVQYYMEHHGQLKEKNGDVIEL-KYPLN-
2pna  -LQDAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFH-RDGKYGFSDPL-TFNSVVELINHYRNES-LAQYNPKLDVKL-LYPVS-
1bfi  HHDEKTWNVGSSNRNKAENLLRGK--RDGTFLVRESS--KQGCYACSVVVDGEVKHCVINKTATG-YGFAEPYNLYSSLKELVLHYQHTS-LVQHNDSLNVTL-AYPVYA
```

(B) DIALIGN multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKT-NGKFLIRAR-DN--NGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKK-FDTLWQLVEHYSYKA-------DGLLRVLT-VPCQK
1gri  EMKPHPWFFGKIPRAKAEEML--SKQRHDGAFLIRESESA--PGDFSLSVKFGNDVQHFKVLRDGAGKYFLWVV-K-FNSLNELVDYHRST--SVSRNQQIFLRDIEQVPQQ-
1aya  M---RRWFHPNITGVEAENLLLTRGV--DGSFLARPSKSN--PGDFTLSVRRNGAVTHIKIQNTGDYYDLYG-GEK-FATLAELVQYYMEHHGQLKEKNGDV-IELK-YPLN-
2pna  LQDAE-WYWGDISREEVNEKL--RDTA-DGTFLVRDA-STKMHGDYTLTLRKGGNNKLIKIFHRDGKYGFSD-PLT-FNSVVELINHYRNE--SLAQYNPKLDVKLL-YPVS-
1bfi  HHDEKTWNVGSSNRNKAENLL--RGKR-DGTFLVRES-SK--QGCYACSVVVDGEVKHCVINKTATGYGFAE-PYNLYSSLKELVLHYQHT--SLVQHNDSLNVTLA-YPVYA
```

(C) ClustalW multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRARDN--NGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKKFD-TLWQLVEHYSYK------ADGLLRVLTVPCQK
1gri  EMKPHPWFFGKIPRAKAEE-MLSKQRHDGAFLIRESES-APGDFSLSVKFGNDVQHFKVLRDGAGKY-FLWVVKFN-SLNELVDYHRSTS-VSRNQQIFLRDIEQVPQQ
1aya  ---MRRWFHPNITGVEAEN-LLLTRGVDGSFLARPSKS-NPGDFTLSVRRNGAVTHIKIQNT-GDYYDLYGGEKFA-TLAELVQYYMEHHGQLKEKNGDVIELKYPLN-
2pna  -LQDAEWYWGDISREEVN--EKLRDTADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFHR-DGKYGFSDPLTFN-SVVELINHYRNES-LAQYNPKLDVKLLYPVS-
1bfi  HHDEKTWNVGSSNRNKAE--NLLRGKRDGTFLVRESSK--QGCYACSVVVDGEVKHCVINKT-ATGYGFAEPYNLYSSLKELVLHYQHTS-LVQHNDSLNVTLAYPVYA
```

(D) divide-and-conquer multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRA-RDNN-GSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKK-FDTLWQLVEHY-SY----KADGLLRV-L-TVPCQK
1gri  EMKPHPWFFGKIPRAKAEEMLS-KQRHDGAFLIRE-SESAPGDFSLSVKFGNDVQHFKVLRDGAGK-YFLWVVK-FNSLNELVDYH-RSTSVSRNQQIFLRDIEQVPQQ-
1aya  ---MRRWFHPNITGVEAENLLL-TRGVDGSFLARP-SKSNPGDFTLSVRRNGAVTHIKIQNTGDYY-DLYGGEK-FATLAELVQYYMEHHGQLKEKNGDVIEL-KYPLN-
2pna  -LQDAEWYWGDISREEVNEKL--RDTADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFHRDGKY-GFSDPLT-FNSVVELINHY-RNESLAQYNPKLDVKL-LYPVS-
1bfi  HHDEKTWNVGSSNRNKAENLL--RGKRDGTFLVRE-SSKQ-GCYACSVVVDGEVKHCVINKTATGY-GFAEPYNLYSSLKELVLHY-QHTSLVQHNDSLNVTL-AYPVYA
```

Alignment can be improved by incorporating additional information

- The alignment of two or more sequences can be improved by incorporating expert knowledge such as known structural properties of one or more sequences.
- For example, if the structure of one of the proteins to be aligned is known, then the gap penalty can be increased for regions of known secondary structures such as a-helices or b-strands, as these regions are less likely to suffer insertions or deletions.
- This will mean that few or no gaps are introduced into these regions.
- On the other hand, gap penalties can be decreased for loop regions, in which insertions and deletions are better tolerated.
- Often the results of an automatic alignment program benefit from manual final adjustment.

Searching Databases

- Searching sequence databases now has a part to play in nearly every branch of molecular biology, and is crucial for making sense of the sequence data becoming available from the genome projects.
- For example, one may wish to search the database with a DNA sequence to locate and identify a gene in a new genome.
- When a protein sequence is available, then searching through the database can be used to identify the potential function.
- Sometimes one wishes to find the gene for a particular protein in a genome, which can be done by searching with a homologous protein or DNA sequence.
- When searching a database with a newly determined DNA or protein query sequence, one does not usually know whether an expected similarity might span the entire query sequence or just part of it.
- Similarly, one does not know if the match will extend along the full length of a database sequence or only part of it.
- Therefore, one initially needs to look for local alignments between the query sequence and any sequence in the database.
- The top-scoring database sequences are then candidates for further analysis.

Search Algorithms

- Fast yet accurate search algorithms have been developed.
- The sequence databases are now extremely large and growing daily.
- This means that aligning a query sequence with sequences in a database requires considerable computer resources.
- In the past, this exceeded the available computing power and so great effort was put into developing fast yet accurate alignment methods.
- Almost all database search programs currently in use are modifications of the rigorous methods discussed earlier.
- The Needleman–Wunsch and Smith–Waterman methods are rigorous in the sense that given a scoring scheme they are guaranteed to find the best-scoring alignments between two sequences.
- Two suites of programs are commonly used for database searching: FASTA and BLAST.

FASTA

- FASTA is a fast database-search method based on matching short identical segments.
- It is a popular database-searching program that increases the speed of a search at the expense of some sensitivity.
- It speeds up the searching process by using k-tuples, short stretches of k contiguous residues.
- In protein searches k can equal 1 or 2, while 6 is a typical value for DNA.
- The program makes up a dictionary of all possible k-tuples within the query sequence.
- Each entry contains a list of numbers that describe the location of the k-tuple in the query sequence, called hashing.
- Therefore, for each k-tuple in the searched sequences, FASTA only has to consult the dictionary to find out if it occurs in the query sequence.

FASTA …

- In the first step of the FASTA method all possible pairwise k-tuples are identified.
- These can be considered as diagonals in a set of dot-plots.
- In the second step, alignments of these diagonals are rescored using a scoring matrix described above.
- In this step, the k-tuple regions are also extended without including gaps, and only those that score above a given threshold are retained.
- In the third step, the program checks to see if some of the highest-scoring diagonals can be joined together.
- Finally, the search sequences with the highest scores are aligned to the query sequence using dynamic programming.
- The final alignment score ranks the database entries and the highest-scoring set is reported.

BLAST

- BLAST (Basic Local Alignment Search Tool) or Wu-BLAST (a version of BLAST developed at Washington University, St Louis) is one of the most widely used database-search program suites.

- It relies on finding core similarity, which is defined by a window of preset size (called a “word”) with a certain minimum density of matches (for DNA) or with an amino-acid similarity score above a given threshold (for proteins).

- Note that these amino acid word-matches do not only include identities and that they are scored with a standard substitution matrix.

- In the first step, all suitable matches are located in each database sequence.

- Subsequently, matches are extended without including gaps, and on this basis the database sequences are ranked.

- The highest-scoring sequences are then subjected to dynamic programming to obtain the final alignments and scores.

- LAST and Wu-BLAST can be run with or without the use of gaps.

Different Versions of BLAST and FASTA

- Many of the search algorithms can be used to search either nucleic acid or protein sequences, or even to search a protein-sequence database using a nucleic acid sequence and vice versa.

- However, you need to choose the correct program for the required type of search.

- In BLAST, for example, one can choose among.
  - ***blastp,*** which compares an amino acid query sequence against a protein-sequence database;
  - ***blastn,*** which compares a nucleotide query sequence against a nucleic acid sequence database;
  - ***blastx,*** which compares a nucleotide query sequence translated in all reading frames against a protein-sequence database;
  - ***tblastn,*** which compares a protein query sequence against a nucleotide-sequence database dynamically translated in all reading frames;

    and finally,
  - ***tblastx,*** which compares the six possible translations of a nucleotide query sequence against the six frame translations of a nucleotide-sequence database.

- The FASTA suite has similar versions of these search programs.

SSEARCH

- Despite the computational requirements, some programs have been written that use rigorous methods to search databases.

- SSEARCH is a search program based on the Smith–Waterman algorithm and is therefore slower than either BLAST or FASTA.

- SSEARCH performs a rigorous search for similarity between a query sequence and the database.

- Other search algorithms based on the Smith–Waterman method have been written and are gaining in popularity as computer power increases.