- Definition: Alignment is the procedure of comparing two or more sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. Alignments may be

1). One to One.

2). One to database.

3). Many to many.

- Origin of sequence similarity are:

1). Homology– Homology is a qualitative term that describes a relationship between genes. Homology implies that the compared sequences diverged in evolution from a common origin.

2). Similarity in function

3). Chance.

- Why do we want alignment?

1). Assigning functions to unknown protein.

2). Determine relatedness of organisms.

3). Identify structurally and functionally important elements.

- Sequence Alignment's general goal:

1). Find maximum degree of similarity. Similarity is a quantitative term that defines the degree of sequence match between two compared sequences.

2). Find minimum evolutionary distance.

Examples 1:

```
G A A C A A T
| | | | | | |
G A A C A A T
```

7/7 or 100%.

Example 2:

```
G A A C A A T
| | |   | | |
G A A T A A T
```

6/7 or 86%.

Types of Sequence Alignments

1). Global vs local alignments.

2). Gapped vs ungapped alignments.

3). Pairwise vs Multiple alignments.

Global Alignments

a). Global Alignment - full sequence alignment.

b). Match as many characters as possible from end to end.

c). Find an alignment with highest total score.

d). Regions of high local similarity may be ignored in favor of a higher overall score.

Local Alignment: –Alignment of subsequence.

a). Find subsequences with highest density of matches.

b). Find regions with high local scores.

c). Sequence similarities may extend beyond the local subsequence with a lower degree of similarity.
Example:

```
- - L A L I G N M - -
    | | | | | |
- - E A L I G N E - -
```

- Gapped vs Ungapped Alignment:
  gap: – A gap is a space introduced into an alignment to compensate for the insertions and deletions in one sequence relative to another.

1. How many gaps need to be inserted into each sequence.

2. Where they need to be placed.

3. How long they must be for optimal alignment.

Example:

A	G	G	V	L	I	I	Q	V	G
|	|	|	|	|	|	×	×	×	×
A	G	G	V	L	I	Q	V	G	-

A	G	G	V	L	I	I	Q	V	G
|	|	|	|	|	×	|	|	|	|
A	G	G	V	L	-	I	Q	V	G

A	G	G	V	L	I	I	Q	V	G
|	|	|	|	|	|	×	|	|	|
A	G	G	V	L	I	-	Q	V	G

To answer the above questions computer must have a scoring function that specifies the quality of the match. This scoring function takes account both matched residues or bases and the gaps:

$$\text{Align quality} = \text{Sum(Score)}_m + \text{Sum(Score)}_g$$

where m represents matches and g represents gaps.

- Score_{m} is typically a similarity measure which is positive for residues that are similar and negative for residues that are different.

- Scores_{g} is negative.

1. Addition of gaps to optimize an alignment always decreases the quality of an alignment.

2. The gap cost is usually defined by the affine gap cost model which simply means that the negative score for a gap as a linear function of the length of the gap, and that the function has a non zero intercept:

$$\text{Score}_{\text{g}} = G_{\text{init}} + G_{\text{extend}} \times \text{length}$$

3. G_{init} can be thought of as the cost of the inserting the first blank character across from a base or residue in a sequence

4. G_{extend} can be thought of as the cost of gap extension or the cost of inserting each additional blank character in an already initiated gap.

- An affine gap penalty encourages the extension of gaps rather than the introduction of new gaps.

- The gap penalties determine the tradeoff between allowing a bad match, i.e., one with a negative score, and inserting a gap, therefore these parameters critically influence the resulting alignment.

- If the gap penalty is too low, a very large numbers of gaps are inserted and almost all sequence can be aligned with proportion of identical bases or residues.

- If the gap penalty is too high, no gaps can be inserted, and the alignment is equivalent to simply sliding the sequences past each other until a maximum score is reached.

  Since there is no way a priori to determine the best gap penalties, it is essential to try a variety of values and examine the affect on the resulting alignment.

- Most alignment programs will suggest gap penalties that are appropriate for a given scoring matrix in most situations.

Similarity vs Distance:

Two ways for measuring homology between sequences. Given two aligned sequences:

- Similarity is a quantitative term that defines the degree of sequence match between two compared sequences. Check how similar they are by counting their matches.

- Check how distant they are by counting their mismatches and indels.

- High similarity is equivalent to low distance.

- Similarity can be either positive or negative.

- Distance is always non-negative.

- Identical sequences have zero distance.

- Pairwise vs Multiple Alignment:
  Pairwise alignments– require 2 sequences.
  Multiple alignments– Require more than two sequences– Computational problem is a lot more.
  Pairwise alignments are fundamental and useful, but when using sequence searching programs (FASTA, BLAST) which perform pairwise alignments to find similar sequences in a database, many sequences are obtained significantly similar to the query sequence. Comparing each and every sequence to every other may be possible when one has just a few sequences, but it becomes impractical as the number of sequences increases.

What we need is multiple sequence alignment, where all similar sequences can be compared in one single figure or table. The basic idea is that the sequences are aligned on top of each other, so that a coordinate system is set up, where each row is the sequence for one protein, and each column is the 'same' position in each sequence.

As with pairwise alignment, there will be gaps in some sequences, most often shown by the dash '-' or dot '.' character. Note that to construct a multiple alignment, one may have to introduce gaps in sequences at positions where there were no gaps in the corresponding pairwise alignment. This means that multiple alignments typically contain more gaps than any given pair of aligned sequences.

Pairwise Alignment Methods

Dot matrix analysis

Dot matrix analysis is based on Dot plots which provide a simple, yet extremely powerful means of nucleic acid sequence analysis. Most commonly they are examined to identify one of two situations:

Regions of similarity within a single sequence (i.e. repeats) or between different sequences. A matrix comparison of two sequences (or one with itself) is prepared by "sliding" a window of user-defined size along both sequences. If the two sequences within that window match with a precision set by the mismatch limit, a dot is placed in the middle of the window signifying a match.

The number of positions taken at a time is the window and the fraction of matches wanted is the stringency. These choices are arbitrary and come with trial and experience.

Example, window of size 5 and stringency of 3. It means up to 2 of 5 bases may mismatch and the 5 base window will still be classified as a match;

window			A	T	C	G	G	C	A	T
1	|	A								
2	||	T								
3	|||	C			$*$					
4	||||	G				$*$				
	||||	G					$*$			
	|||	A						$*$		
	||	G								
	|	T								

when mismatch limit is set to 0, all 5 bases must match perfectly in order for the windows to match.

We can adjust the "stringency" of the match by adjusting window size and mismatch limit - the large the window of comparison, and the lower the mismatch limit, the most stringent the comparison.

Dynamic Programming algorithm

It is a technique or algorithm which, when implemented correctly, effectively makes all possible pairwise comparisons between the characters (nucleotide or amino acid residues) in two biological sequences. The final result is a mathematically, but not necessarily biologically, optimal alignment of the two sequences. A similarity score is also generated to describe how similar the two sequences are, given the specific parameters used.

- Needleman -Wunsch —— for global alignment.

  This algorithm guarantees the best scoring for alignment between two sequences.
  Input two sequences
  $S = s_1 s_2 \ldots s_n$
  $T = t_1 t_2 \ldots t_m$
  n and m are approximately same.

- Complexity of Needleman Wunsch Algorithm

1). Time complexity — need to fill all cell tables $O(n \times m)$

2). Space Complexity — Similarity $O(n \times m)$.

3). Two sequences of length 1000 symbols require 1,000,000 operation.

- Smith Waterman —— for local a alignment.

- Same complexity problem as in Needleman Wunsch .

- Heuristic Alignment
  A procedure that progresses along empirical lines by using rules of thumb to reach a solution i.e., based on what is experienced or seen rather than on theory . The solution is not guaranteed to be optimal.

1). Advantage —— Very fast —— reliable in a statistical sense.

2). Disadvantage —— Less sensitive than Dynamic programming

- Two heuristic Methods — Fasta & Blast.

1). Fasta: Comparing query sequence against a database of sequences (1985).

2). Blast: Basic local alignment search technique improvement of Fasta.

- Common idea for a good alignment contains subsequences of absolute identity.

1). First, identify very short (almost) exact matches.

2). Next, the best short hits from the Ist step are extended to longer regions of similarity.

3). Finally, the best hits are optimized using the Smith Waterman algorithm.

Introduction of GCG SeqWeb

SeqWeb is a new web interface to a subset of the various protein and nucleic acid analysis programs found in the Wisconsin Package (GCG). In our computer lab this package is available under the web browser:
http://darwin.sci.csuhayward.edu
Each student needs his/her own login and password to get in the package.