Multiple Sequence Alignment Introduction

What is an alignment?

- An arrangement of two or more DNA, RNA or protein sequences
  - A multiple sequence alignment is one of more than two sequences
- Homologous sites between sequences are aligned
  - This is achieved by inserting gaps

Why do we align?

- Alignments allow for identification of regions of similarity between sequences

- Identify indels (insertion and deletions) caused by DNA/RNA replication

What is an alignment used for?

- Building phylogenetic trees

- Looking for sites of interest/conservation within a gene (motifs, binding sites, etc.)

- Identifying positive/negative selection

- Using as references for short read analysis

How do we align?

- The goal of alignment programs is to maximise a score based on 3 modifiers:
  - rewarding matches (+ score)
  - penalising rare substitutions (- score)
    - requires a substitution matrix
  - penalising gaps (- score)
    - requires gap opening and extension penalty scheme
- Different programs go about it through different methods

Substitution matrices

- Places weights upon comparison of characters
- Most simple is +1 for a match and 0 for a mismatch
- As some substitutions are more acceptable than others these must be weighted
- Substitution matrices assign scores to each substitution
  - Built from alignments of given similarity
    - BLOSUMx where x is the similarity
    - PAMx where x is the number of subsitutions/100 amino acids

BLOSUM62

Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

Gaps

- Represent an insertion in 1 or more sequences or a deletion in the remaining sequences
- Two types of penalties are associated with gaps:
  - Gap opening
    - To make an initial opening of a gap in a sequence
  - Gap extension
    - To add an extra gap character to an existing gap
    - Gap extension penalty usually smaller than gap opening

Pairwise alignment

- Dot-matrix
  - 1 sequence as a row, 1 sequence as a column
  - A dot where two characters match
- Dynamic programming
  - Use a scoring function to optimally align sequences
  - Either a global or local algorithm used
- Word
  - Heuristic method using ‘words’ of a given size
  - Find matching words and extend alignments until 1 sequences ends or score drops below a threshold

Global versus local

- A global alignment method attempts to align sequences end-to-end
  - Useful when sequences are of approximately the same length
  - Needleman-Wunsch algorithm
- A local alignment method attempts to find one or more stretches of similar sequences
  - Useful when one sequence is significantly longer than the other or there are small similar motifs within large dissimilar sequences
  - Smith-Waterman algorithm

Multiple Sequence Alignment

- Progressive
  - Do a pairwise alignment
  - Use a clustering method to create a guide tree
  - Using the guide tree create a succession of pairwise alignments starting with the two closest sequences and ending with the most distant from these
- Iterative
  - Given a MSA remove a sequence and realign to the others
  - May also optimise weights and distance measures
  - Repeat to convergence

MUSCLE

- A progressive alignment is created starting with a word-based pairwise method
- A new distance matrix is created from this
- The old and new trees are compared and sequences realigned to reflect new guide tree
  - If old and new tree are the same we stop
- The alignment is split into 2 profiles and these are aligned as above
  - Different bipartitions are tried until convergence is reached

MAAFT

- Options for a standard progressive alignment (word based)
- Iterative alignment available using guide tree reconstruction and realignment
- Can use dynamic programming (local or global) instead of word based initial pairwise alignment

Editing alignments

- Trimming
  - Removal of poorly aligned regions can improve subsequent analysis
  - Cut-offs of gap proportions or amino acid variation (entropy) are used to remove columns
- Manual
  - Some sequences are difficult for optimal automated aligning
  - Manual editing of alignments based on users biological knowledge may improve alignments