Multiple Sequence Alignment Lab Introduction

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
  - Done by programs such as BMGE or Gblocks
- Manual
  - Some sequences are difficult for optimal automated aligning
  - Manual editing of alignments based on users biological knowledge may improve alignments