multiple sequence alignment

workshop on molecular evolution, north america 2011
expected learning outcomes

- basic alignment types
- affine gap penalties
- canonical progressive alignment algorithm
- improvements to the canonical progressive alignment algorithm
- alignment challenges
- name recognition of several alignment programs
sequence alignment, a definition

the arrangement of nucleotides or amino acids from two or more sequences, accounting for insertions and deletions where necessary, such that nucleotides or amino acids at each position are considered homologous

thus, the resulting alignment is a compound statement of implied positional homology
types of alignments

• based on the number of sequences
  - pairwise: two sequences are aligned; the fundamental step in sequence database searching (e.g., with BLAST or FASTA)
  - multiple sequence alignment: three or more sequences

• based on the portion of the sequences
  - global: alignment is over the entire length of the sequences (e.g., Needleman-Wunsch algorithm)
  - local: alignment of segments of all lengths to find that which is optimal (e.g, Smith-Waterman algorithm, FASTA, BLAST)
scoring function

• match/mismatch function
  - nucleotides: may differentially weight transitions and transversions
  - amino acids: reflect some notion of similarity, often based on substitution matrices such as the BLOSUM and PAM series

• gap penalty function
  - constant gap penalty
  - linear gap penalty
  - affine gap penalty
affine gap penalty

• combines the parameter of constant gap penalty with that of linear gap penalty

• formally defined as, \( \text{gap penalty} = a + bx \)

  \( a \) is the gap opening penalty, the penalty for starting a gap; the only parameter of a constant gap penalty

  \( b \) is the gap extension penalty, the penalty for each position in the gap; the only parameter of a linear gap penalty

  \( x \) is the length of the gap, number of positions

• dependent on the number and length of gaps
canonical progressive alignment algorithm

1. generate pairwise alignments
2. generate a distance matrix
3. generate a distance-based guide tree (e.g., UPGMA, NJ)
4. align sequence groups based on relationships in the guide trees starting with most closely related and progressing through most distantly related

the underlying concept is that the more closely related sequences are more easily and accurately aligned

important reference - Feng and Doolittle 1987 *J Mol Evol* 60:351-360
improving progressive alignments

• differential weighting of end gaps
• iterative alignment
  - use alignment to generate new distance matrix
  - use new distance matrix to make new guide tree
  - use new guide tree to generate new alignment
• refinement
  - change position of nucleotides/amino acids/gaps within one or more sequences
• iterative refinement
• new objective functions
  - weighted sum-of-pairs (WSP)
  - consistency between multiple and pairwise alignments
• different algorithms for different alignment problems
progressive method (A) and iterative refinement method (B)
potential multiple sequence alignment problems

- directionality of gap placement
  - gap placement differs when the order of the residues is reversed
  - heads or tails approach: compare regular and reversed order alignments to identify problem regions
- once a gap always a gap
- overlapping gaps
- data may fit neither the local nor global alignment model
  - partial length and full length gene sequences (end gap problem)
  - cDNA and genomic DNA
  - alternatively spliced genes or gene products
global (A), local (B), and long internal gaps (C)
some specific newer ideas
phylogeny-aware gap placement

• for progressive alignment there is an asymmetry in how insertions and deletions are handled
  - bias toward fewer inferred insertions, more inferred deletions
  - resulting alignments are more compact

• a few recent and relevant papers -
fast statistical alignment (FSA)

- based on pair hidden Markov models, which approximate an insertion/deletion process on a tree
- uses a sequence annealing algorithm to combine the posterior probabilities estimated from models into a multiple alignment
- produces estimates of the alignment accuracy and uncertainty for every column and character of the alignment
simultaneous alignment and phylogenetic inference

• alignment and phylogenetic inference are best not treated as independent analytical steps, first (popularly) articulated by Felsenstein (1988)
• ALIFRITZ, BAi-Phy, POY, POY*, SATCHMO are programs that seek to achieve simultaneous alignment and phylogenetic inference
• more recently: SATé (simultaneous alignment and tree estimation)
MAFFT (multiple alignment fast Fourier transform)

- demonstrably the best, or near best, performing multiple sequence alignment program in numerous objective tests
- developed by Katoh and colleagues (see publications from 2002-2010) and under continuing development
fast group-to-group alignment algorithm based on FFT

A Convert an amino acid sequence to a 2D wave

LAFAKTNVK → Volume

B Convert a profile to a 2D wave

WYDAREAAL V---ADRAGV → Volume

C Correlation coefficient

\[ c(k) \]

D Restrict the area of the DP matrix

sequence 1

sequence 2
profile, group-to-group, and skeleton alignment
reading list (starting points)


Katoh K. MAFFT web site, particularly algorithms and tips (other sections are useful as well)

   http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html