MCB 5472 Lecture #6:
Sequence alignment
March 27, 2014

Sequence alignment

- As you have seen, sequence alignment is key to nearly all experiments in molecular evolution
- Thus far we have discussed local alignment as implemented in BLAST
- Global alignment:
  - Aligns sequences over their entire length
  - Assumes that sequences for alignment are homologous

Recall from BLAST lecture:

- Sequence alignment is scored:
  - According to a substitution matrix
    - Some substitutions are more likely than others
  - Using affine gaps
    - Gap opening and extension are considered separately
    - Reflects biological reality that
- Alignment score is the sum of substitution and gaps scores

Pairwise alignments

Needleman-Wunsch

- Needleman and Wunsch (1970) J. Mol. Biol. 48:443-453
- The first algorithm to computer the optimum alignment between two sequences using dynamic programming
  - i.e., examines many possible solutions and picks the best
- Implemented as EMBOSS `needle` program
- Global alignments: assumes sequences should be aligned over their entire lengths

Needleman-Wunsch

- Works by scoring alignments sequentially and evaluating scoring for each alignment position based on previous scores
  - Gaps in one position may force an unlikely substitution later on
  - Calculating these scores represents a dynamic programming sub-problem; computationally efficient
- Evaluates all possibilities but also maps the best option
  - Guarantees finding the best path according to the parameters used

Smith-Waterman

- Smith and Waterman (1981) J. Mol. Biol. 147:195-197
- Local alignment version of Needleman-Wunsch
- Guaranteed to find the statistically best local alignment
  - BLAST only evaluates a subset of alignment possibilities
  - Implimented in FASTA search program

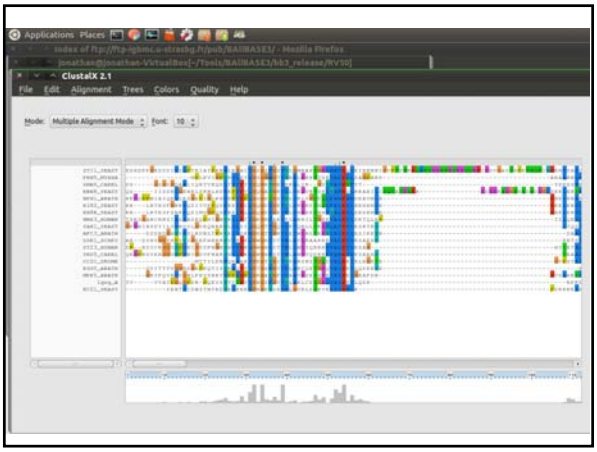
Multiple sequence alignment

Multiple sequence alignment

- Extending similar dynamic programming approaches to calculate all possible sequence alignments quickly becomes impossible
- Various tools therefore use different heuristic approaches to align multiple sequences
  - Different specializations and/or motivations
  - Different computational efficiency

ClustalW

- One of the first widely used multiple sequence alignment programs
- Thompson et al. (1994) Nuc. Acids Res. 22: 4673-4680
- Larkin et al. (2007) Bioinformatics 23:2947-2948
- ClustalX: Widely used version with a graphical interface



ClustalW

- Step #1a: Align all pairs of sequences separately
  - Current default: count kmers conserved between sequences
  - Can also be global alignments (original defaults)
- Step #1b: Calculate distance matrix from pairwise comparisons
- Step #2: Cluster distance matrix using neighbor-joining or UPGMA algorithms to create a “guide tree”
- Step #3: Midpoint root tree and weight branches by sequence similarity

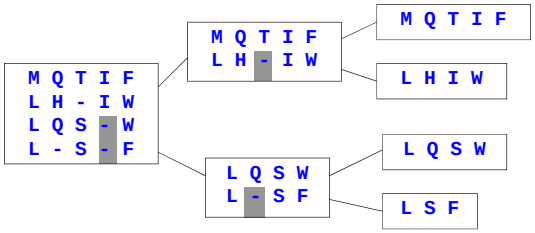
ClustalW guide trees

- Guide trees are no substitute for full phylogenetic analysis!!!
  - Not based on multiple sequence alignment
- Are only a rough approximation of the true relationships between sequences
- Even though they can be produced by ClustalW they should not be used for detailed analysis!

ClustalW

- Step #4: Progressive alignment
  - Starting from most similar sequences on guide tree, align each to each other
  - Uses full dynamic programming alignment methods (cf. N-W) including substitution and gap penalties
    - Gap opening parameters vary based on sequence position to favor alignment to preexisting gaps
  - Any gaps introduced are maintained during subsequent alignment iterations

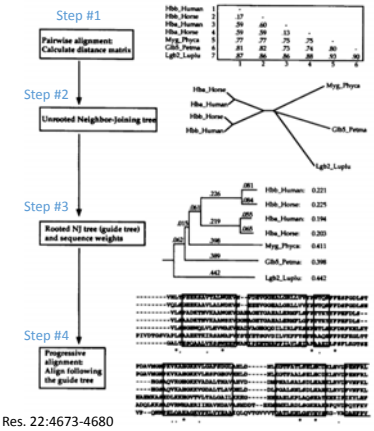
Progressive alignment example



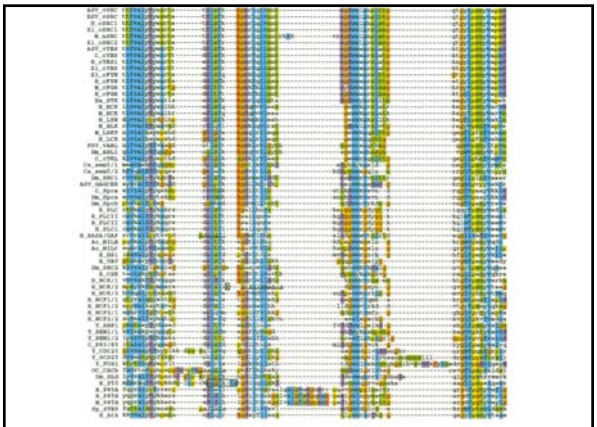
- Gaps introduced earlier are propagated into later alignments

Edgar (2004) BMC Bioinformatics 5:133

ClustalW



Thompson et al. (1994) Nuc. Acids Res. 22:4673-4680

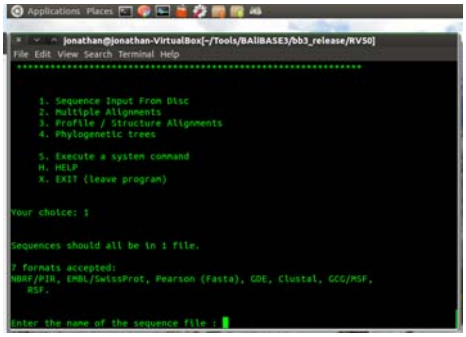


Thompson et al. (1994) Nuc. Acids Res. 22:4673-4680

ClustalW

- No guarantee of optimal alignment
  - Early errors propagated to more divergent sequences
  - This is true of all multiple sequence programs
- Reasonably accurate when all sequences are ~ <40% identical
- Scales reasonably well to a few thousand sequences
- Easy to run!
- Has been superseded by better programs

ClustalW command line



Easy! Type “clustalw” and follow along

MUSCLE

- Edgar (2004) Nuc. Acids Res. 32:1792-1797
- Edgar (2004) BMC Bioinformatics 5:133
- Designed to improve speed and accuracy over older multiple sequence alignment programs like clustalw
- Now a preferred alignment method, especially for high-throughput studies

MUSCLE

- Stage #1: create draft progressive alignment
  - Step #1-1: calculate distance between genomes using kmers, create distance matrix
  - Step #1-2: cluster distance matrix into guide tree using UPGMA algorithm
  - Step #1-3: conduct progressive multiple sequence alignment
    - Accuracy sacrificed for speed at this step
- So far, same as clustalw but faster and less accurate

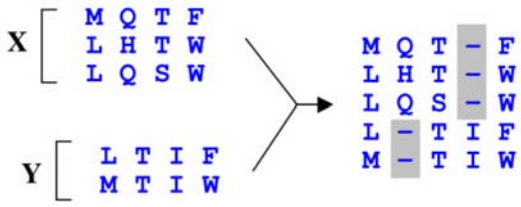
MUSCLE

- Stage #2: refine progressive alignment
  - Most inaccuracy in Stage #1 due to using kmer distances to create guide tree
  - Step #2-1: re-create distance matrix using Kimura distances (more accurate, need input multiple sequence alignment)
  - Step #2-2: cluster distance matrix using UPGMA
  - Step #2-3: recalculate progressive alignment, omitting alignments that stayed the same from Step #1-3
- Result: more accurate alignment than Stage #1

MUSCLE

- Stage #3: Alignment refinement
  - Step #3-1: moving from root to tip in the tree from step #2-2, remove that node and spit the alignment into two subalignments
  - Step #3-2: compute alignment profiles for each subalignment
  - Step #3-3: re-align profiles to each other
  - Step #3-4: determine if new alignment has a better score than the previous one, if so keep new one and goto step #3-1 using the next node in the tree
    - Stop when scores stop improving

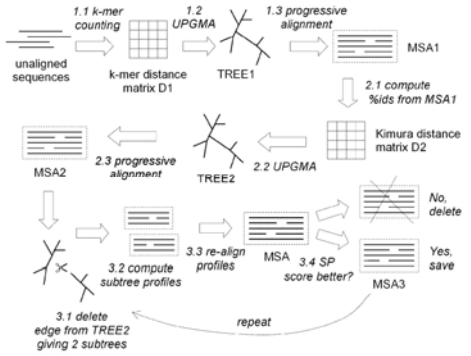
Profile alignment



- Can yield better gap placement by removing biases from original pairwise alignments

Edgar (2004) BMC Bioinformatics 5:133

MUSCLE



Edgar (2004) Nuc. Acids Res. 32:1792-1797

MAFFT

- Katoh et al. (2002) Nucl. Acids Res. 30:3059-3066
- Katoh & Standley (2013) Mol. Biol. Evol. 30:772-780
- Similar to MUSCLE, designed to improve speed and accuracy of multiple sequence alignment vs. clustalw

MAFFT

- Instead of progressive alignments, MAFFT uses a "fast Fourier transform"
  - Creates local alignment blocks based on physico-chemical properties of amino acids (esp. volume & polarity)
  - Very fast!
  - Does not require calculating alignments exhaustively, rather how blocks link together
- Statistical framework the same for pairwise and multiple alignments

MAFFT

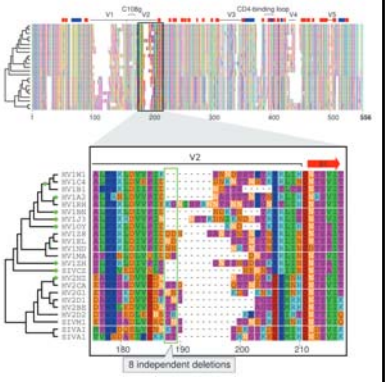
- Scoring system dramatically simplified relative to clustalw (uses complicated heuristic normalizations)
- Contains an optional iterative refinement method (similar to MUSCLE)
- Newer versions contain robust profile alignment methods (i.e., aligning alignments)
- Speedup and accuracy similar to MUSCLE

PRANK

- Loytynoja and Goldman (2008) Science 320:1632-1635
- Motivation: alignment programs typically group gaps together
  - Gaps represent insertion/deletion (indel) evolutionary events
- Result: multiple evolutionary events are grouped together

clustalw

- clustalw SIV gp120 protein alignment
- Reconstruction of indels implies 8 independent deletions
  - (unlikely)

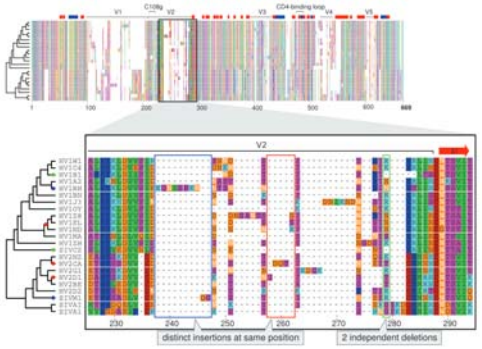


Loytynoja and Goldman (2008) Science 320:1632-1635

PRANK

- Progressive alignments using substitution matrices sequentially evaluates alignments on a column-by-column basis
- Individual columns by themselves lack sufficient information to accurately reflect evolution of the entire sequence
- PRANK evaluates gap conservation during alignment refinement to decide if the gap should be used during subsequent alignment steps

PRANK of same SIV gp120



Loytynoja and Goldman (2008) Science 320:1632-1635

PRANK

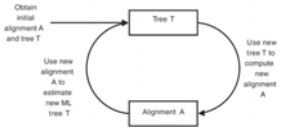
- Better models of indel events
- Sequence alignments are not artificially compressed (i.e., shorter than true alignments)
- Computational cost

SATé

- Liu et al. (2012) Syst. Biol. 61:90-106
- Liu et al. (2009) Science 324:1561-1564
- Something different: performs alignment and tree estimation simultaneously

SATé

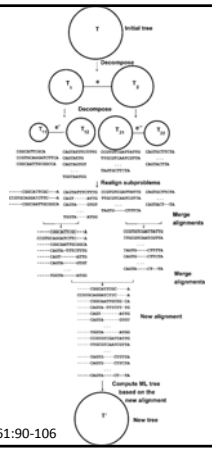
- Multiple iterations creating new trees and alignments each time
- Trees constructed using Multiple Likelihood methods (v. robust)
- ML trees function as “guide” trees for subsequent iterations



Liu et al. (2009) Science 324:1561-1564

SATé

- Alignments subdivided along longest tree branch each iteration
  - Many splits having increasing phylogenetic resolution
- Alignments merged into master alignment
- Tree recalculated



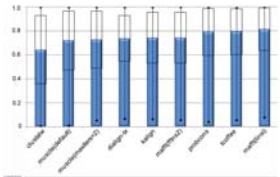
Liu et al. (2012) Syst. Biol. 61:90-106

Which alignment method is best?

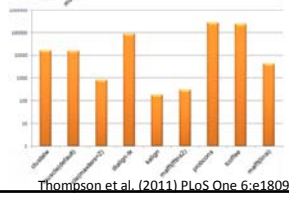
- Head-to-head analyses rarely cover all possibilities
- Depends on the expected output
  - E.g., comparison to reference alignment
  - E.g., effect on tree construction
  - E.g., effect on identifying site-specific selection
- Trade-offs: speed vs accuracy

Which alignment method is best?

Alignment score compared to known reference

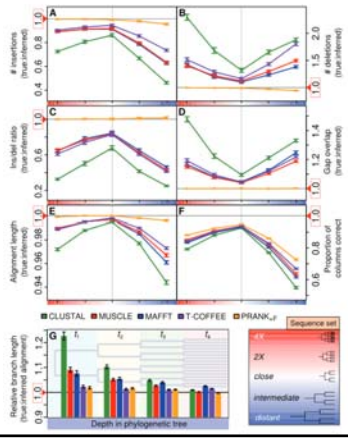


Compute time



Thompson et al. (2011) PLoS One 6:e18093

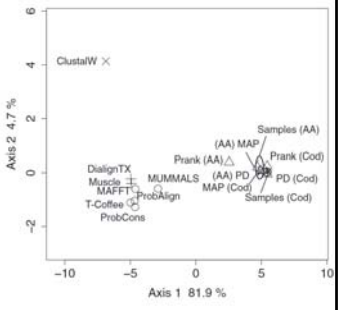
Which alignment method is best?



Loytynoja and Goldman (2008) Science 320:1632-1635

Different algorithms can give different results

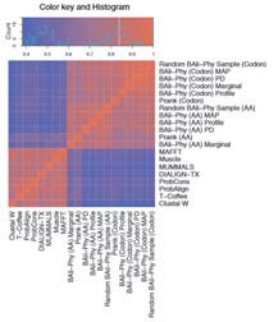
- PCoA plot of trees constructed using different alignment algorithms



Blackburne and Whelan (2013) Mol. Biol. Evol. 30:642-653

Different algorithms can give different results

- Correlation of sites identified as under selection using different sequence alignment algorithms



Blackburne and Whelan (2013) Mol. Biol. Evol. 30:642-653

Note: nucleotides vs. proteins

- Recall: proteins are more conserved compared to nucleotides
- Sequence alignment is therefore more robust using protein sequences
- Gaps in nucleotide alignments should reflect codon structures

Note: nucleotides vs. proteins

- Software exists to convert between protein and nucleotide sequences for alignment
- PAL2NAL http://www.bork.embl.de/pal2nal/ Suyama et al. (2006) Nucl. Acids Res. 34:W609-W612
- MEGA6: GUI version http://www.megasoftware.net/ Tamura et al. (2013) Mol. Biol. Evol. 30:2725-2729
  - Doesn't scale fantastically compared to terminal but user-friendly

Sequence masking

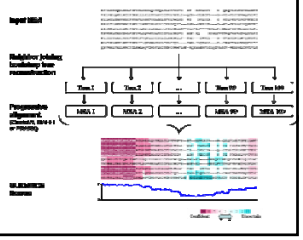
- Another way to deal with poor alignments is to remove regions thought to be inaccurate before further analysis
  - Lose information, but optimize sensitivity/specificity tradeoff
- Common for phylogenetic applications
  - Gaps are often used during phylogenetic reconstruction, so are better to remove if not actually informative
- Not entirely without controversy

Gblocks

- Talavera and Castresana (2007) Syst. Biol. 56: 564-577
- http://molevol.cmima.csic.es/castresana/Gblocks.html
- Identifies blocks of sequences aligned with high confidence
  - E.g., few gaps, few columns lacking sequence conservation, confidently-aligned flanking regions

GUIDANCE

- Penn et al. (2010) Nucl. Acids Res. 38:W23-W28
- http://guidance.tau.ac.il/overview.html
- Create alignment guide trees based on alignment columns
- Score compared to master alignment



Summary:

- Choice of multiple sequence alignment program will affect downstream analyses
- Different trade-offs to approach
- No substitute for manual inspection and correcting alignments when the resulting phylogeny really, really matters!-