

# Practical Bioinformatics

Mark Voorhies

5/31/2013

# Exercise: Scoring a gapped alignment

- 1 Given two equal length gapped sequences (where “-” represents a gap) and a scoring matrix, calculate an alignment score with a -1 penalty for each base aligned to a gap.
- 2 Write a new scoring function with separate penalties for opening a *zero length* gap (e.g.,  $G = -11$ ) and extending an open gap by one base (e.g.,  $E = -1$ ).

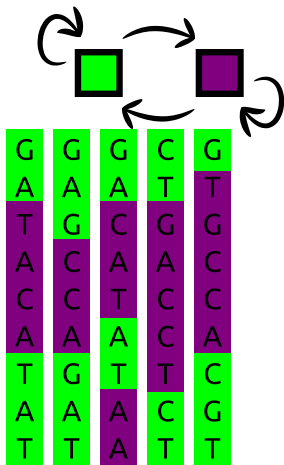
$$S_{gapped}(x, y) = S(x, y) + \sum_i^{\text{gaps}} (G + E \cdot \text{len}(i))$$

# HMMer3 sensitivity and specificity

fraction of

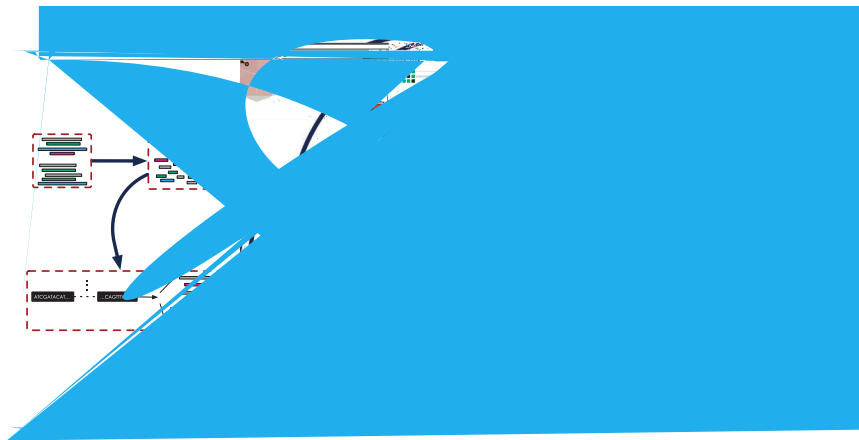
1  
0.9  
0.8  
0.7  
0.6  
0.5  
0.4  
0.3  
0.2  
0.1  
0

# EM: Training an HMM



- If we have a set of sequences with known hidden states (*e.g.*, from experiment), then we can calculate the emission and transition probabilities directly
- Otherwise, they can be iteratively fit to a set of unlabeled sequences that are known to be true matches to the model
- The most common fitting procedure is the Baum-Welch algorithm, a special case of expectation maximization (EM)

# EM: Estimating transcript abundances



Roberts and Pachter, Nature Methods 10:71

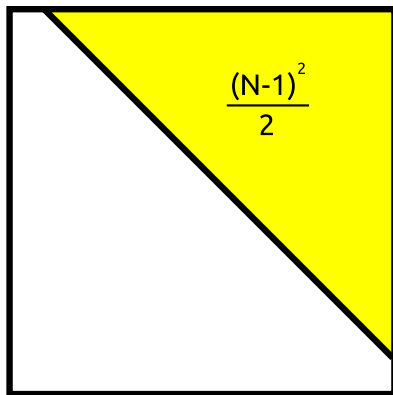
# Evolution implies a self-consistent model



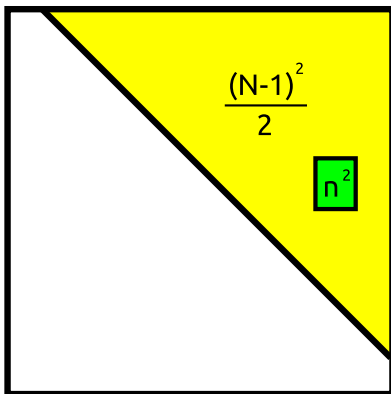
Distances  
(Pairwise relationships)

Topology  
(Evolutionary history)

# Measure all pairwise distances by dynamic programming

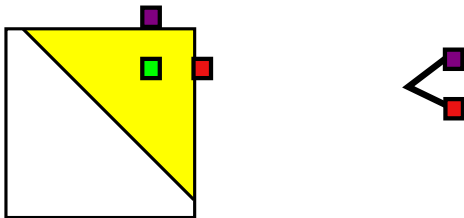


# Measure all pairwise distances by dynamic programming

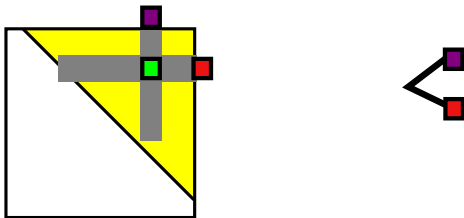




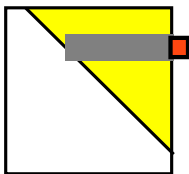
# Generate a guide tree by UPGMA



# Generate a guide tree by UPGMA



# Generate a guide tree by UPGMA

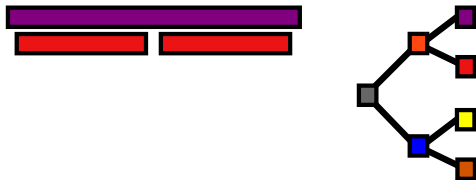


# Generate a guide tree by UPGMA

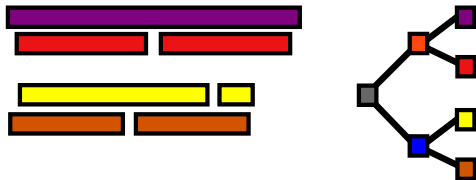




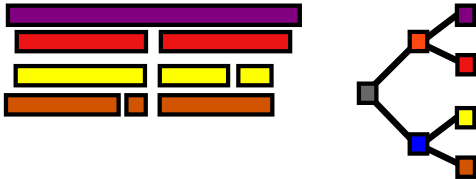
# Progressive alignment following the guide tree



# Progressive alignment following the guide tree

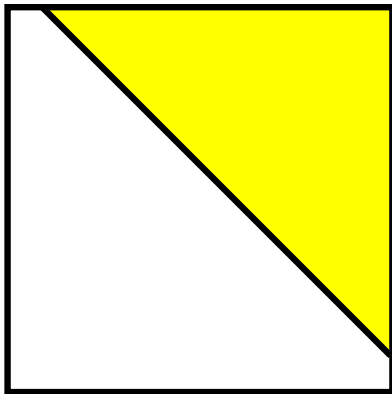


# Progressive alignment following the guide tree

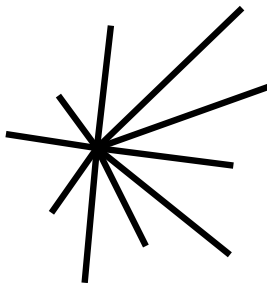
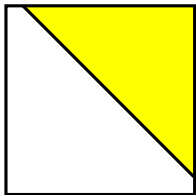




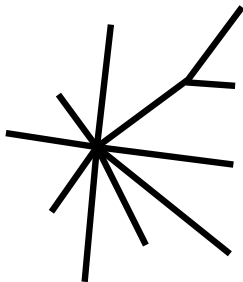
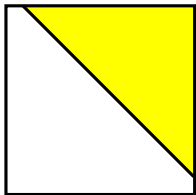
# Measure distances directly from the alignment



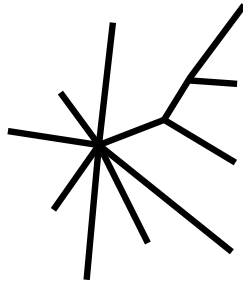
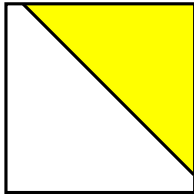
# Generate neighbor-joining tree from new distances



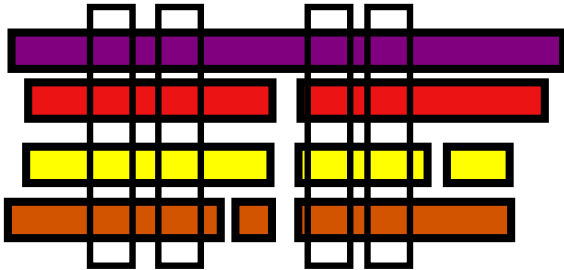
# Generate neighbor-joining tree from new distances



# Generate neighbor-joining tree from new distances



# Generate bootstrap values from subsets of the alignment



# Generating a multiple alignment in CLUSTALX

The screenshot displays the ClustalX 2.0.12 application window. The title bar reads "ClustalX 2.0.12". The menu bar includes "File", "Edit", "Alignment", "Trees", "Colors", "Quality", and "Help". The "File" menu is open, showing options: "Load Sequences" (Ctrl+O), "Append Sequences", "Save Sequences as..." (Ctrl+S), "Load Profile 1", "Load Profile 2", "Save Profile 1 as...", "Save Profile 2 as...", and "Write Alignment as Postscript" (Ctrl+P). A dropdown menu is currently set to "0".

The main window contains a multiple sequence alignment of protein sequences. The sequences are color-coded by amino acid type. The alignment is shown in a grid format with columns representing positions across different sequences. Below the alignment, there are several sections for writing profiles and alignments as Postscript files, including "Write Profile 1 as Postscript" and "Write Profile 2 as Postscript".

At the bottom of the window, there is a status bar with navigation icons (back, forward, search, etc.) and a scale bar.

# Generating a multiple alignment in CLUSTALX

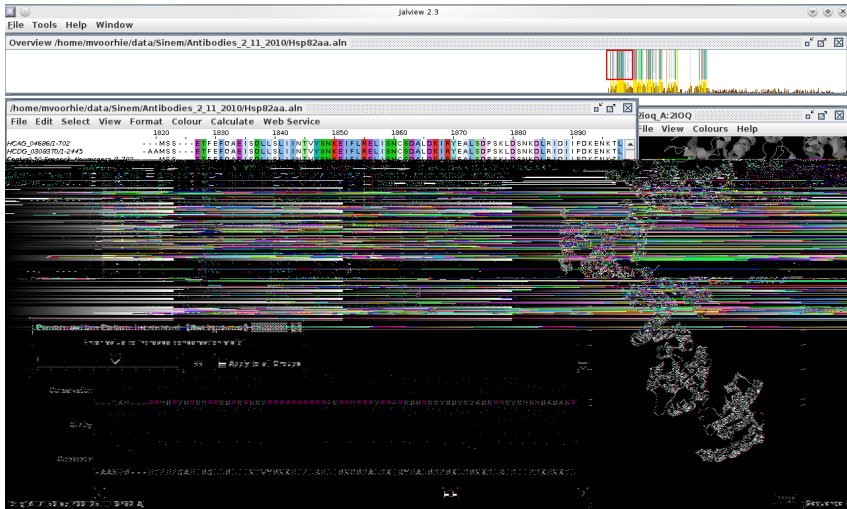
The screenshot displays the ClustalX 2.0.12 software interface. The main window shows a multiple sequence alignment of protein sequences, with each amino acid residue represented by a colored letter. A menu is open over the 'Alignment' tab, listing several options: 'Do Complete Alignment' (Ctrl+L), 'Do Guide Tree Only' (Ctrl+G), 'Do Alignment from Guide Tree', 'Realign Selected Sequences', 'Realign Selected Residue Range', 'Align Profile 2 to Profile 1', 'Align Profiles from Guide Trees', 'Align Sequences to Profile 1', and 'Align Sequences to Profile 1 from Tree'. The background shows a complex alignment view with a dendrogram on the left and a sequence alignment grid on the right. The interface includes a menu bar (File, Edit, Alignment, Trees, Colors, Quality, Help) and a status bar at the bottom.

# Generating a neighbor joining tree in CLUSTALX

The screenshot displays the ClustalX 2.0.12 software interface. The main window is divided into several panels. On the left, a large area shows a multiple sequence alignment of protein sequences, with gaps represented by dashes. Above the alignment, there are various alignment statistics and parameters. On the right, a smaller panel displays a neighbor-joining tree, which is a hierarchical clustering diagram showing the relationships between the sequences. The tree is rooted and shows the branching order of the sequences. The interface includes a menu bar at the top with options like File, Edit, Alignment, Trees, Colors, Quality, and Help. At the bottom, there are navigation buttons and a status bar.



# Viewing the alignment and tree in JALVIEW



- Multiple Alignment
  - T-Coffee
  - MUSCLE
  - COBALT
  - hmalign (HMMer3)
- Tree building
  - MrBayes (Bayesian MCMC)
  - PhyML (maximum likelihood)
  - FastTree2 (very large heuristic trees)

Finish your dynamic programming implementation.