Practical Bioinformatics

Mark Voorhies

5/31/2013

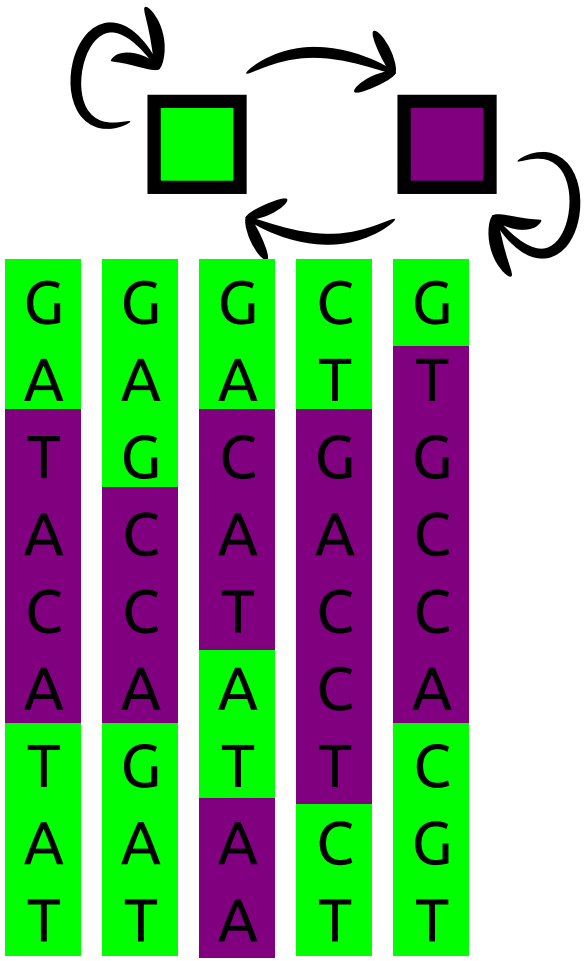
Exercise: Scoring a gapped alignment

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).

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

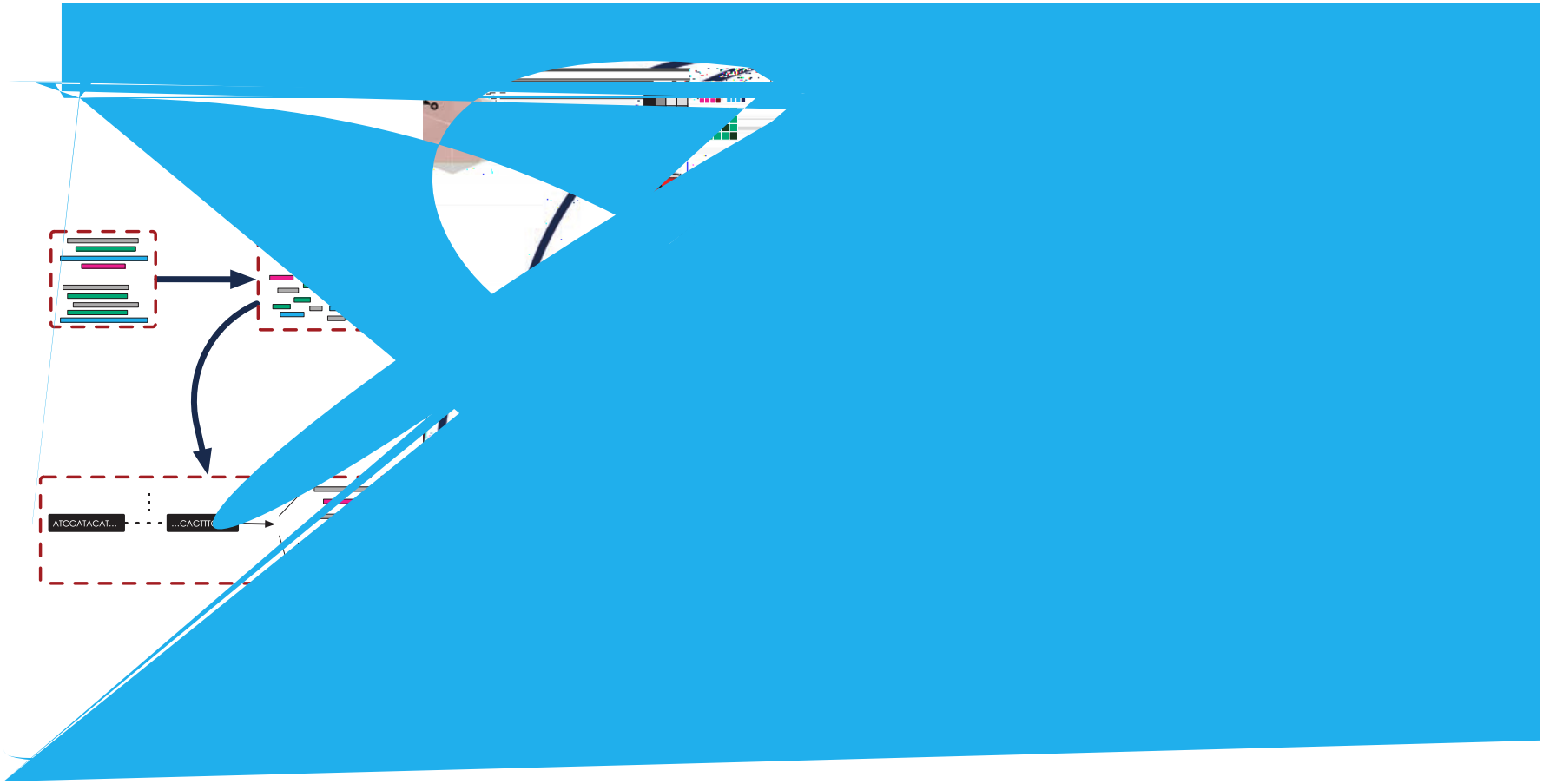
HMMer3 sensitivity and specificity

EM: Training an HMM



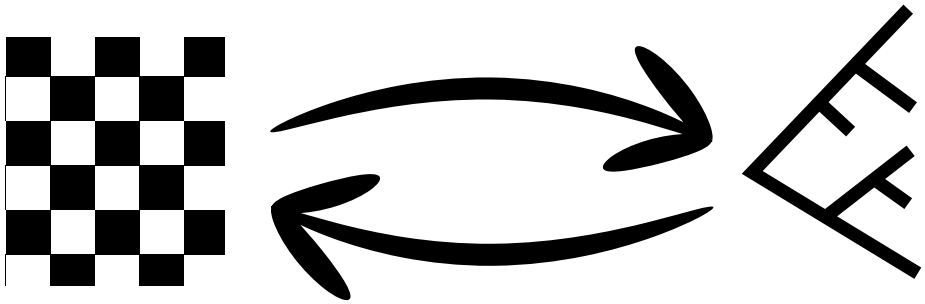
- If we have a set of sequences with known hidden states (*e.g.*, from experiment), then we can calculate the emission and transition probabilities directly
- Otherwise, they can be iteratively fit to a set of unlabeled sequences that are known to be true matches to the model
- The most common fitting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

EM: Estimating transcript abundances



Roberts and Pachter, Nature Methods 10:71

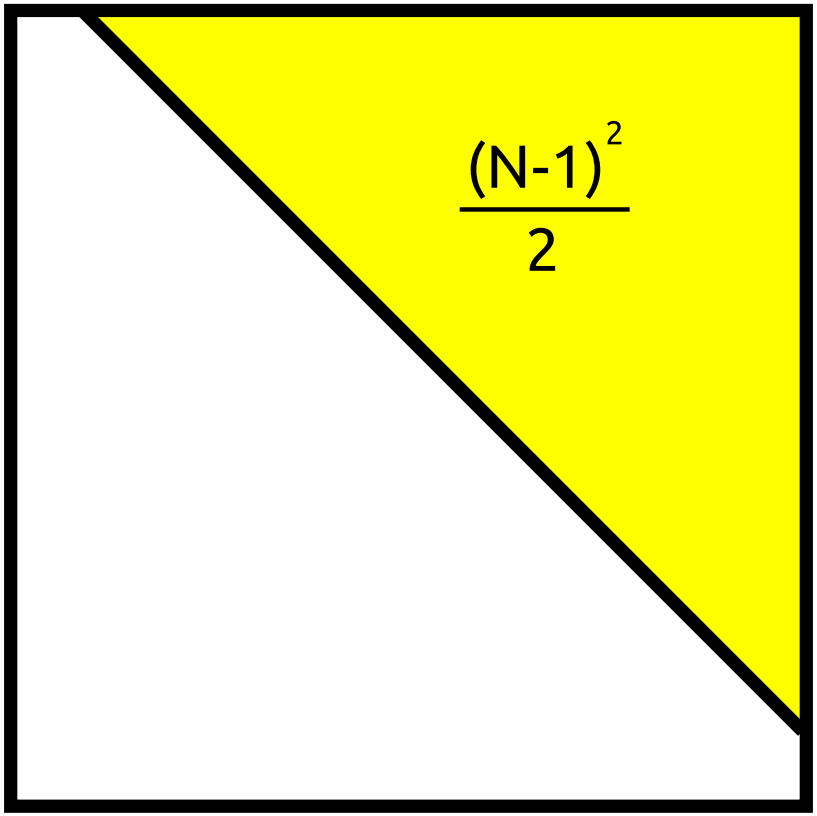
Evolution implies a self-consistent model



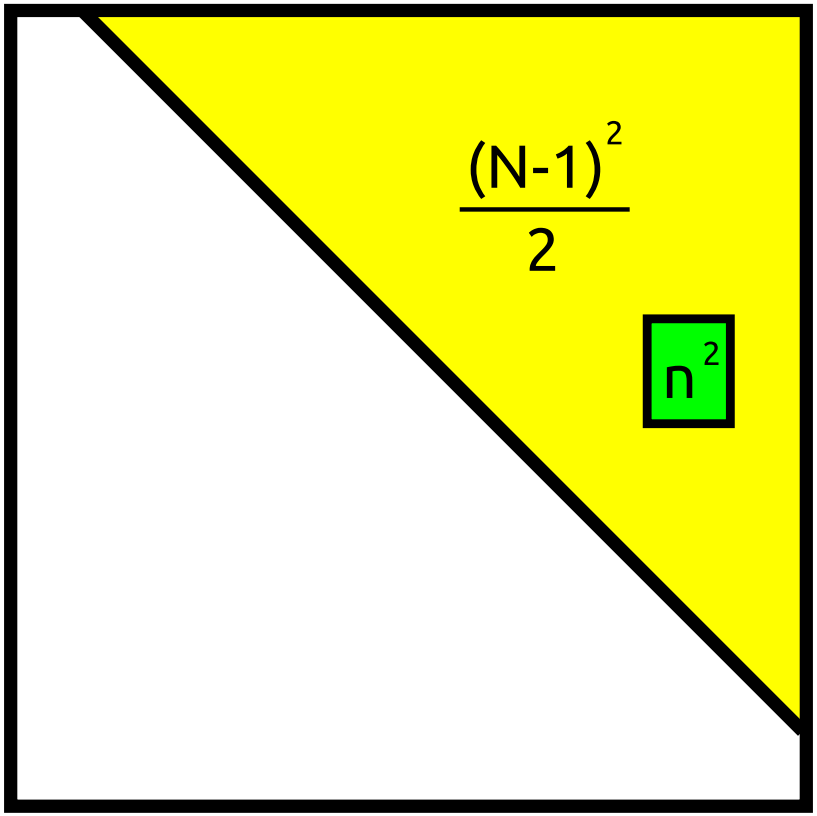
Distances
(Pairwise relationships)

Topology
(Evolutionary history)

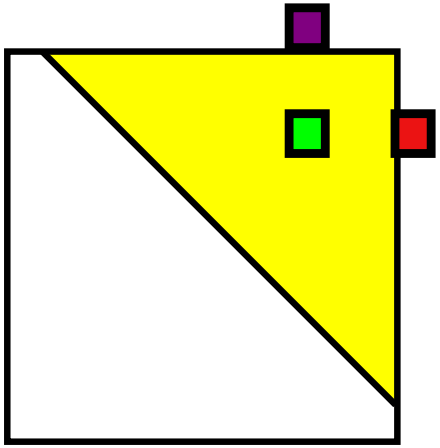
Measure all pairwise distances by dynamic programming

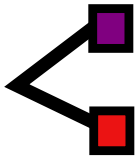


Measure all pairwise distances by dynamic programming

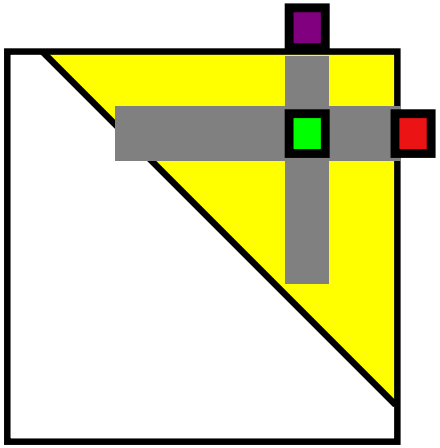


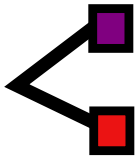
Generate a guide tree by UPGMA



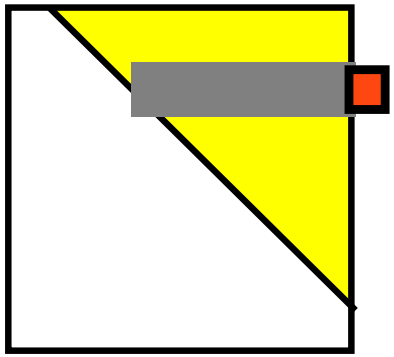


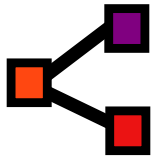
Generate a guide tree by UPGMA



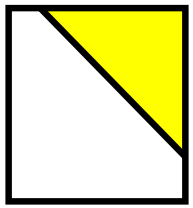


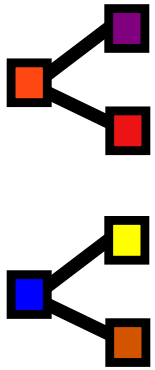
Generate a guide tree by UPGMA



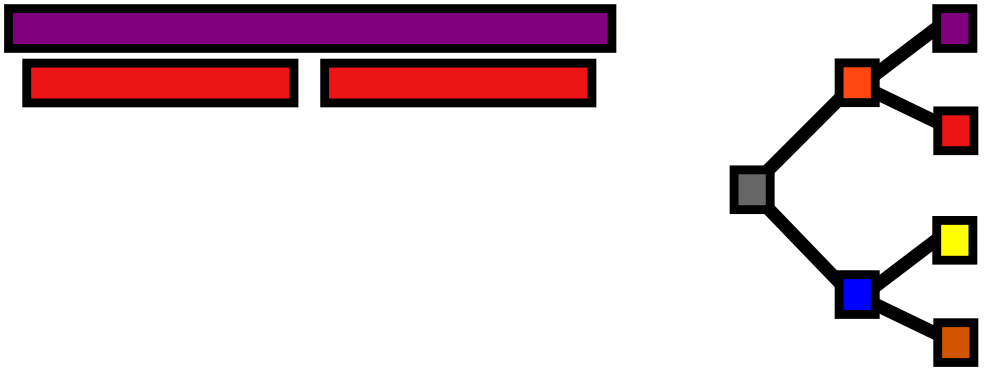


Generate a guide tree by UPGMA

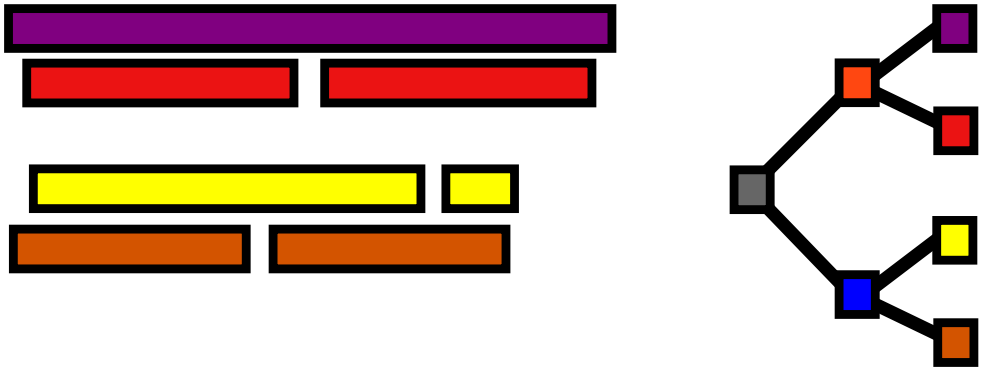




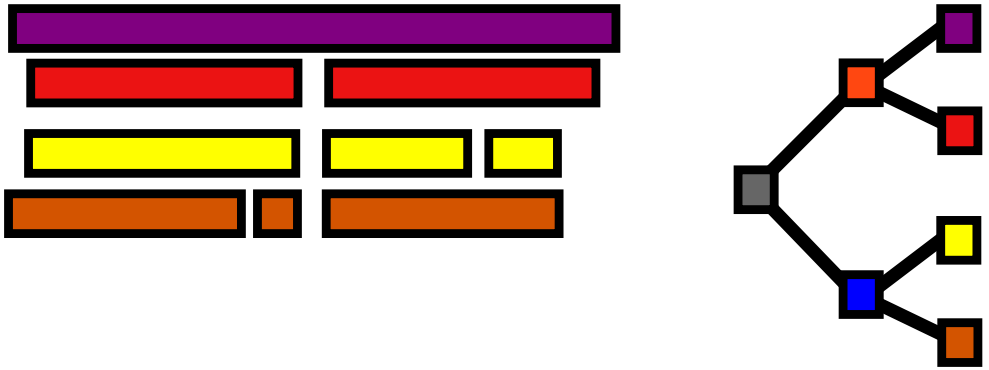
Progressive alignment following the guide tree



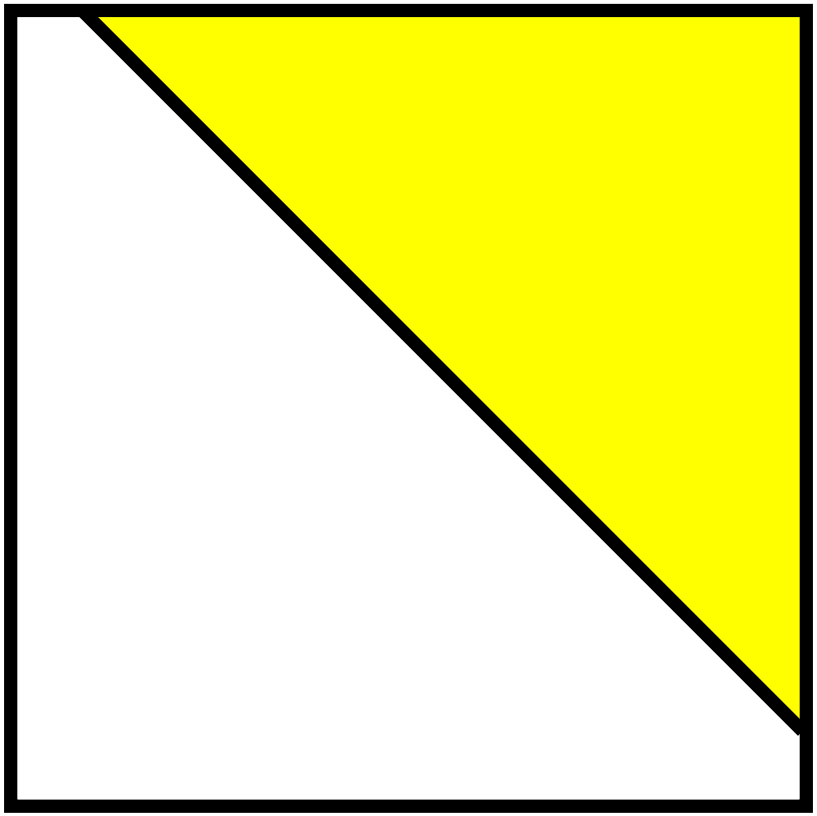
Progressive alignment following the guide tree



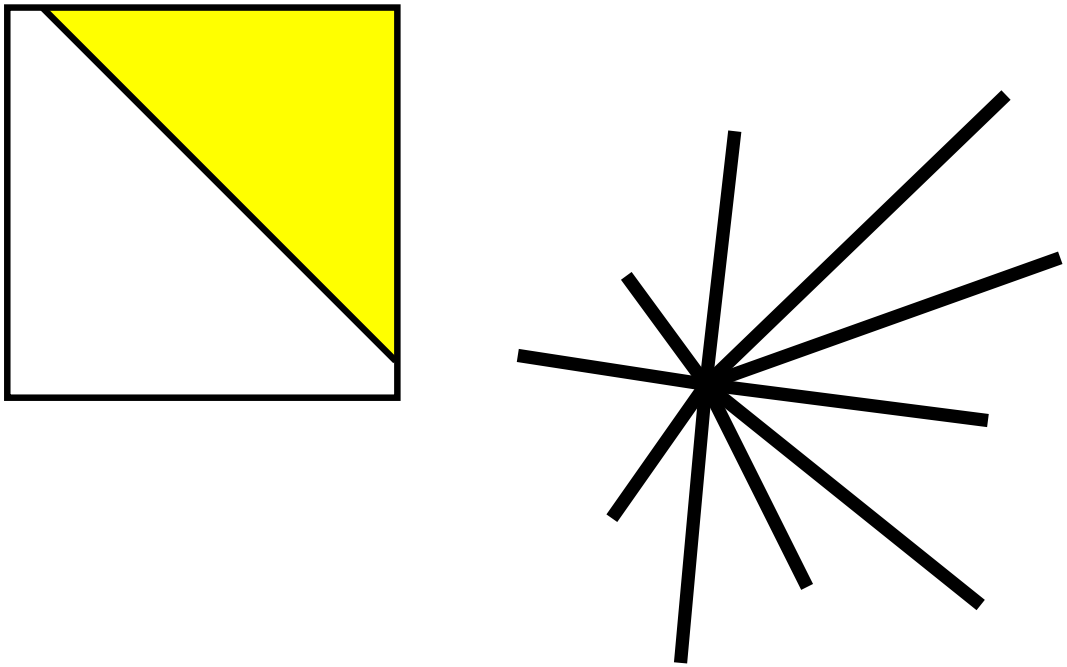
Progressive alignment following the guide tree



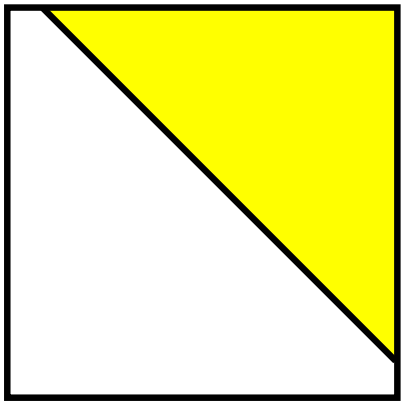
Measure distances directly from the alignment

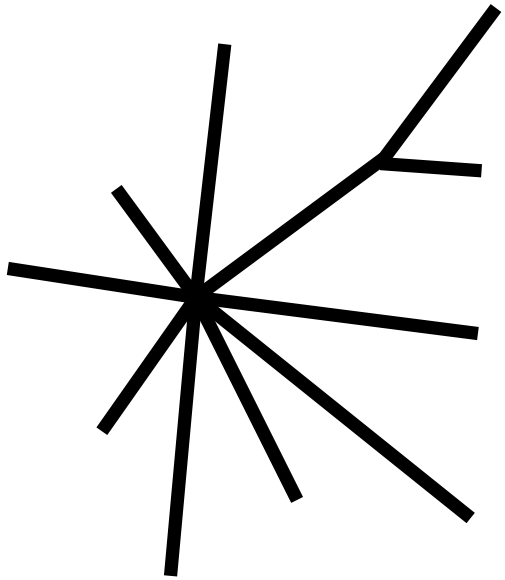


Generate neighbor-joining tree from new distances

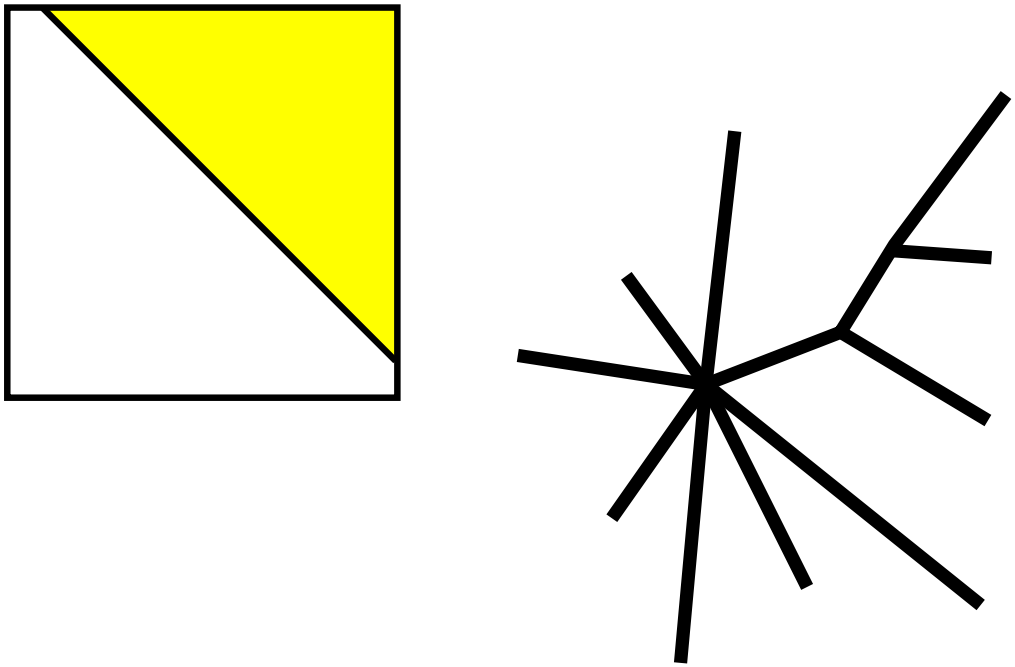


Generate neighbor-joining tree from new distances

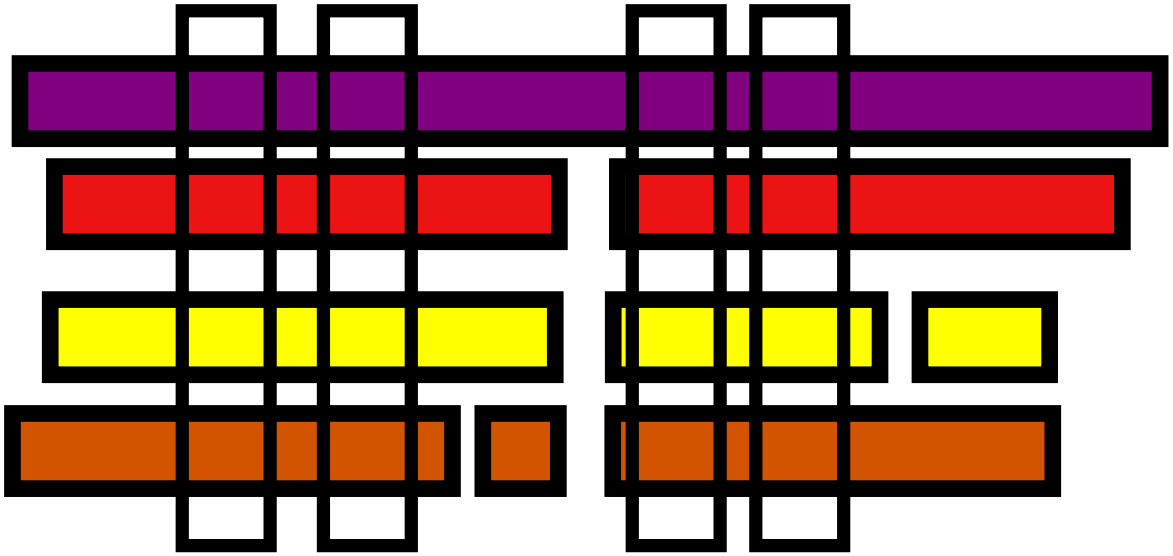




Generate neighbor-joining tree from new distances

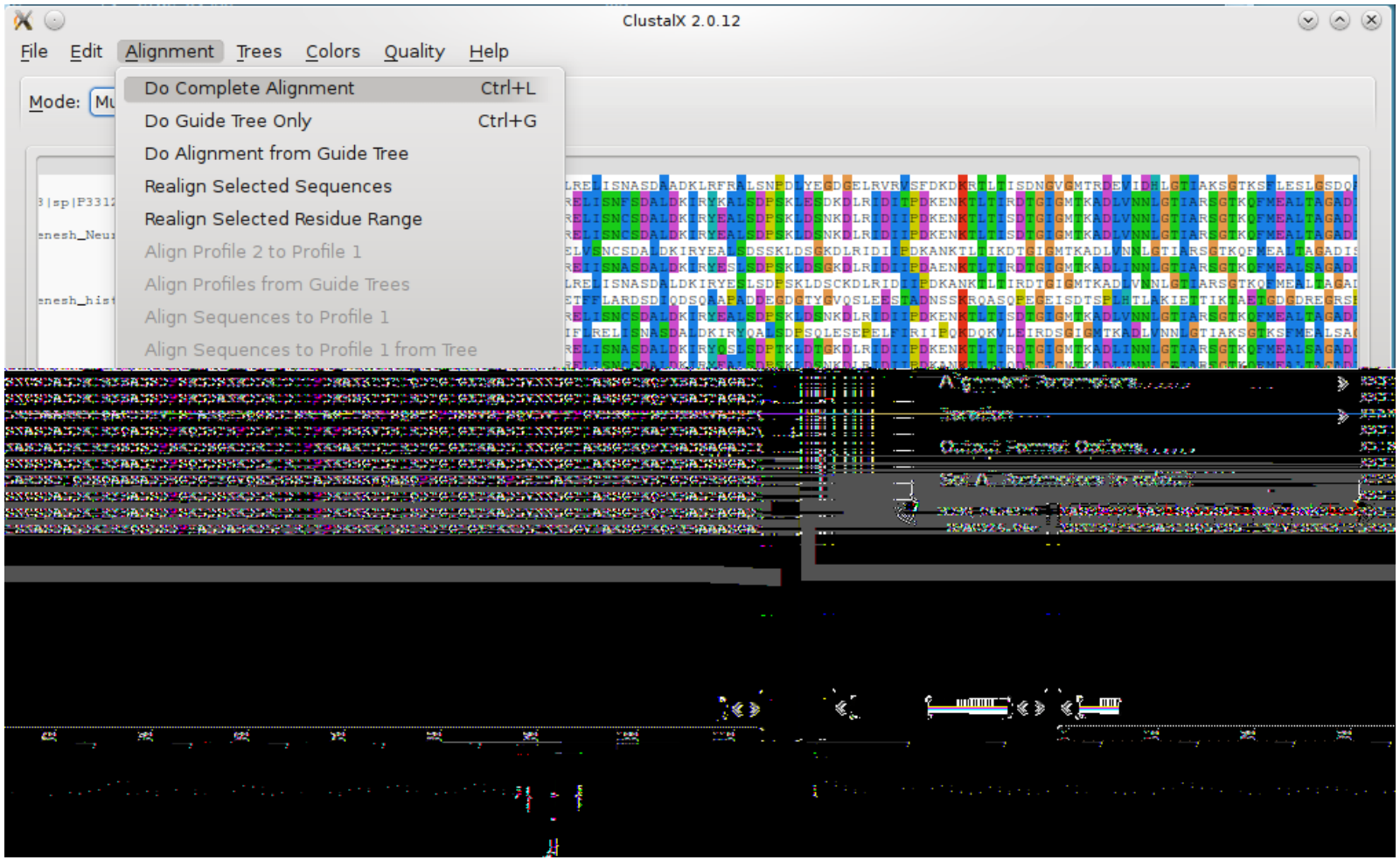


Generate bootstrap values from subsets of the alignment



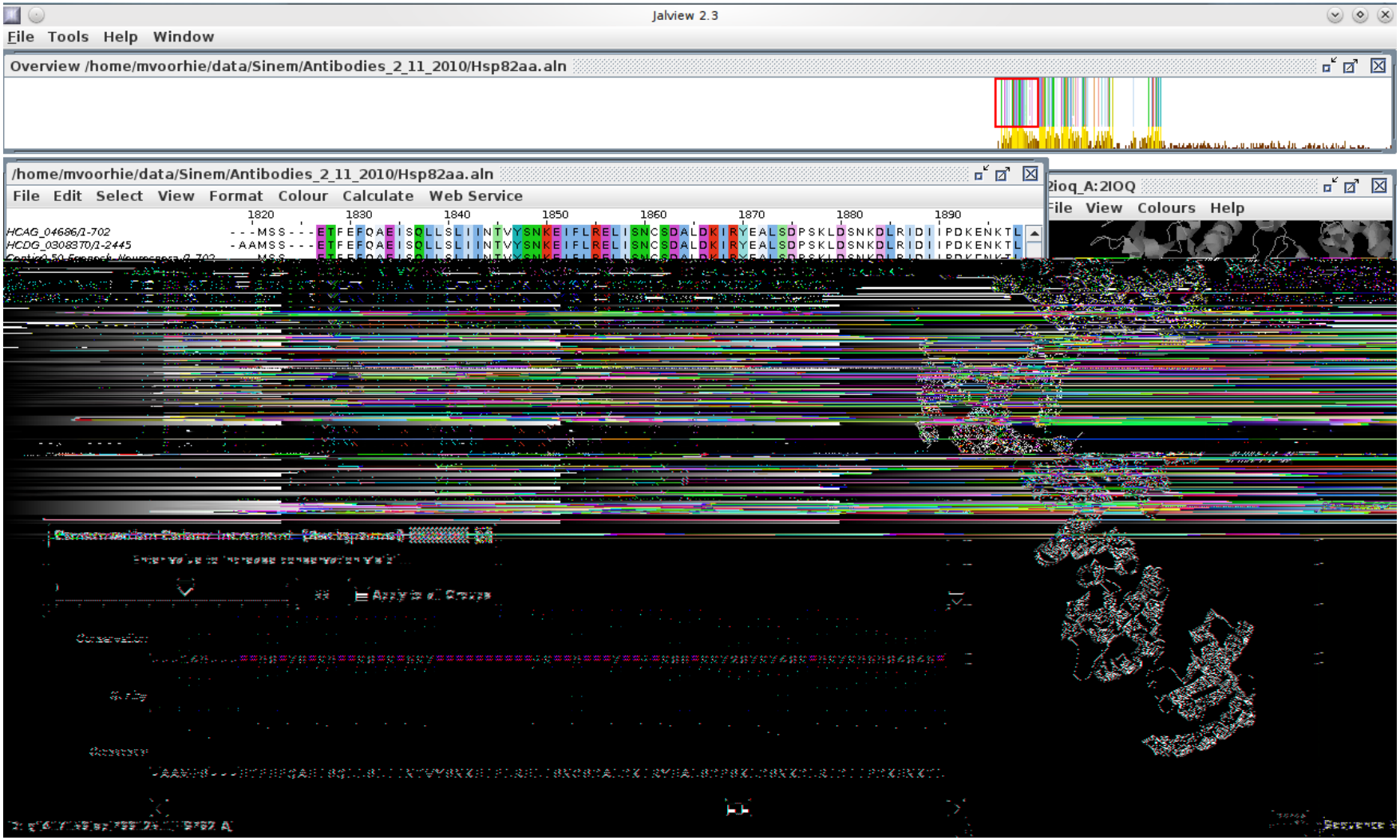
Generating a multiple alignment in CLUSTALX

Generating a multiple alignment in CLUSTALX



Generating a neighbor joining tree in CLUSTALX

Viewing the alignment and tree in JALVIEW



Related tools

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

Homework

Finish your dynamic programming implementation.