

# Practical Bioinformatics

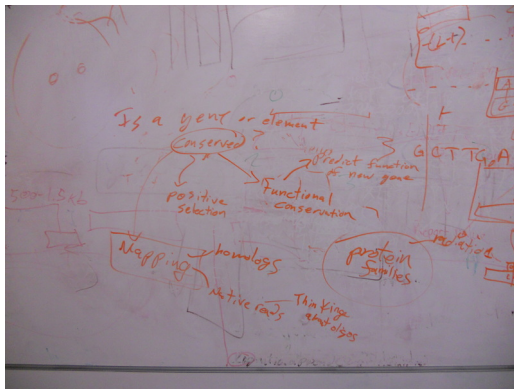
Mark Voorhies

5/26/2015

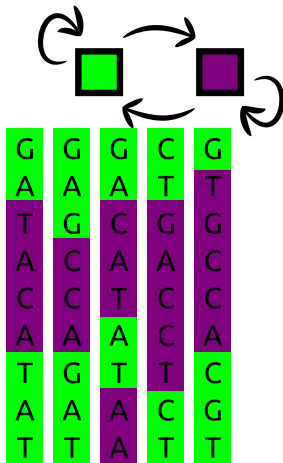
*Habits are things you get for free, without requiring any special work.*

*–Cory Doctorow Advice to Writers, 4/5/2012*

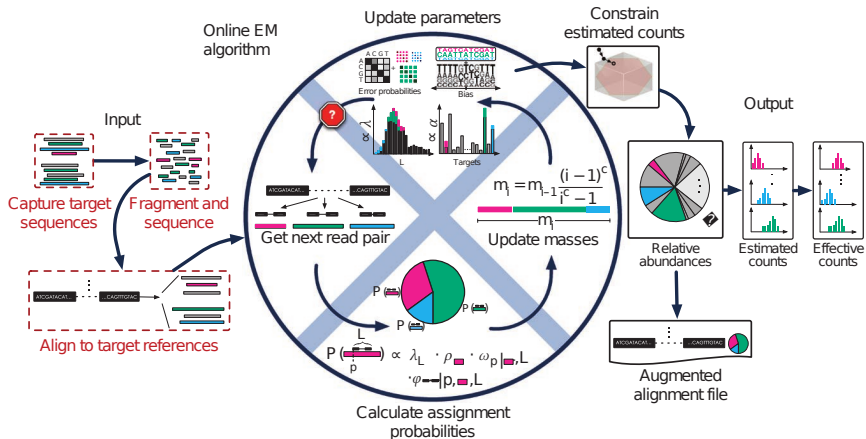
# Why compare sequences?



# EM: Training an HMM



# EM: Estimating transcript abundances



Roberts and Pachter, Nature Methods 10:71

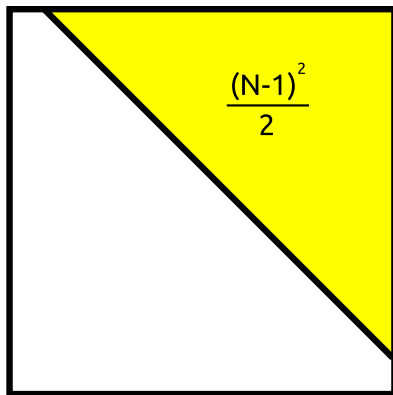
# Evolution implies a self-consistent model



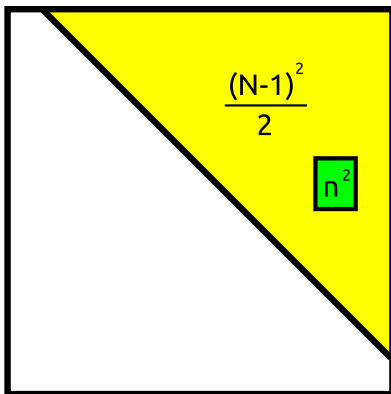
Distances  
(Pairwise relationships)

Topology  
(Evolutionary history)

# Measure all pairwise distances by dynamic programming

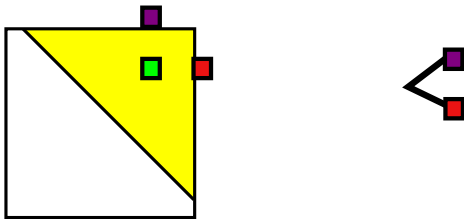


# Measure all pairwise distances by dynamic programming

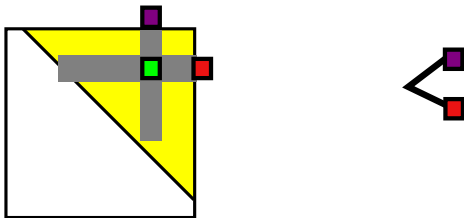




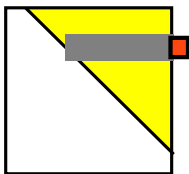
# Generate a guide tree by UPGMA



# Generate a guide tree by UPGMA



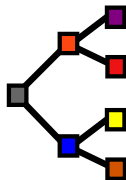
# Generate a guide tree by UPGMA



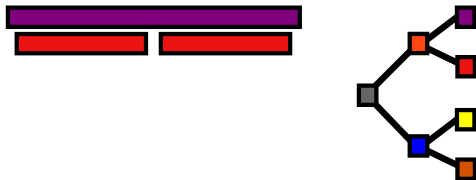
# Generate a guide tree by UPGMA



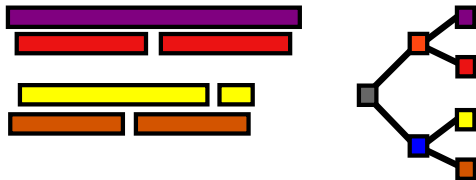
# Generate a guide tree by UPGMA



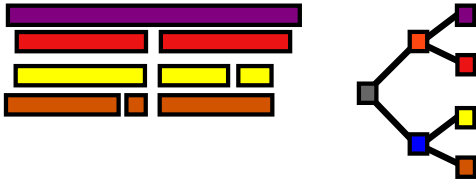
# Progressive alignment following the guide tree



# Progressive alignment following the guide tree

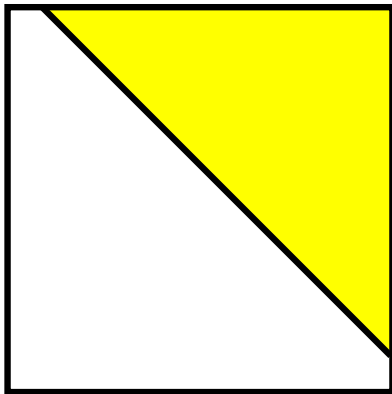


# Progressive alignment following the guide tree

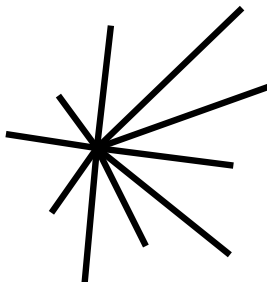
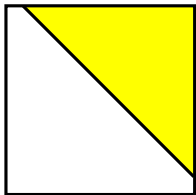




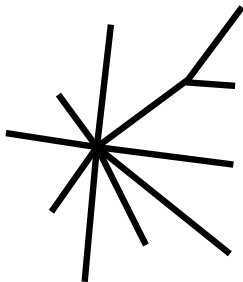
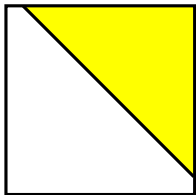
# Measure distances directly from the alignment



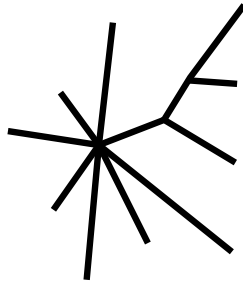
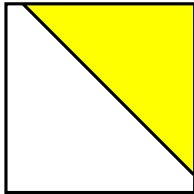
# Generate neighbor-joining tree from new distances



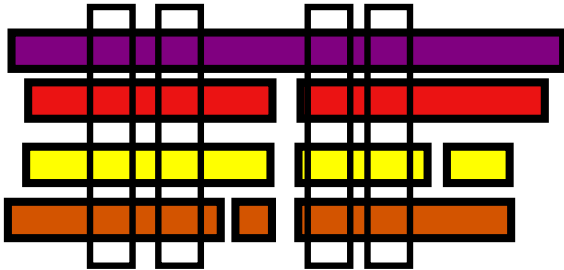
# Generate neighbor-joining tree from new distances



# Generate neighbor-joining tree from new distances



# Generate bootstrap values from subsets of the alignment



# Generating a multiple alignment in CLUSTALX

ClustalX 2.0.12

File Edit Alignment Trees Colors Quality Help

- 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...
- Write Alignment as Postscript Ctrl+P
- Write Profile 1 as Postscript
- Write Profile 2 as Postscript
- Quit Ctrl+Q

0 v

A009010200062D SEYFFFOAEIISD MS NYV SNKPT REL NASDA DKIRYEA DSKLDS KDLRID I NKEAKT I NDT I MTKADI N I L T T S E T K P F S A P A A D T  
\_CMAQ\_04150\_1 L I L R E Y S N K E L R E I S N A S D A D K L R F R A G N D I Y E D E L R V R S P D K G L I S P N V E N T R I L I L I L A K E T K E R E E F S D Q  
BCEE\_031407D L I I I V Y S N K E I F R E L I L G A L D K R E A S E S P D K R I D I D I S K E N S R E I Q E M K R A V I G L I L A R K R E A L A G D I  
PAAQ\_056797D L I I I V Y S N K E I F R E L I L G A L D K R E A S E S P D K R I D I D I S K E N S R E I Q E M K R A V I G L I L A R K R E A L A G D I  
BDBQ\_045437D L I I I V Y S N K E I F R E L I L G A L D K R E A S E S P D K R I D I D I S K E N S R E I Q E M K R A V I G L I L A R K R E A L A G D I  
SPAC92.6.04c S N T E L K F E A E I S D M S I N Y S N K E L R E I S N A S D A D K I R Y E A S D D R L D A E K D I I N T E K E N K L S R D T I M T K N D I N I L A K S E T K P R E A P A A D T

1 10 20 30 40 50 60 70 80 90 100 110

# Generating a multiple alignment in CLUSTALX



# Generating a neighbor joining tree in CLUSTALX

The screenshot displays the CLUSTALX 2.0.12 application window. The 'Trees' menu is open, showing the following options:

- Draw Tree (Ctrl+R)
- Bootstrap N-J Tree (Ctrl+B)**
- Exclude Positions with Gaps
- Correct for Multiple Substitutions
- Output Format Options
- Clustering Algorithm >

The main window shows a multiple sequence alignment of protein sequences. The alignment is color-coded by amino acid type. A progress bar at the bottom indicates the current position in the alignment, ranging from 1550 to 1594. A status bar at the bottom left reads: "CLUSTAL-Alignment file created [Hsp82aa.aln]".



# Viewing the alignment and tree in JALVIEW

The screenshot displays the JALVIEW software interface. The main window shows a multiple sequence alignment of protein sequences from various sources, including *HCAG\_046862.702*, *HCDO\_03083702-2445*, *Contig0\_30\_Fgenseh\_Neurospora\_1-702*, *HCIG\_03340702-1221*, *gi417153|sp|P33125.1|HSPB2\_A1-679*, *INSTD\_FE\_Contig19\_Fgenseh\_insl/2-1614*, *HCDO\_011599702-704*, *BOBO\_04548702-704*, *PADO\_07715702-671*, *PADO\_06249702-290*, *PADO\_05679702-495*, *CIMG\_047292-702*, *Af3g0421702-706*, *HC0901020006201-699*, *AN809.2/2-700*, *JMG\_06759\_5\_2-702*, *JNC00142.2\_1-705*, *gk19\_65252-707*, *YMR186W2-705*, *SPAC206.cnc2-704*, *CIMG\_06159\_1\_1-699*, and *Dirg\_A1-624*. The alignment is color-coded by conservation, with a 'Conservation Colour Increment (Background)' dialog box open, showing a slider set to 55 and the 'Apply to All Groups' option checked. Below the alignment, there are three bar charts: 'Conservation' (yellow), 'Quality' (yellow), and 'Consensus' (black). The sequence ID at the bottom is `Sequence 5 ID: gi417153|sp|P33125.1|HSPB2_A1`. On the right side, a 3D ribbon diagram of a protein structure is shown, with a 'Jmol' label at the bottom right. The ribbon diagram is rendered in a dark grey color with some colored highlights. The Jmol window title is 'Jmol A:2IQ' and it has a menu bar with 'File View Colours Help'.

- Protein Multiple Alignment
  - MUSCLE
  - Clustal Omega
  - Probcons
  - hmalign (HMMer3)
- Tree Building
  - MrBayes (Bayesian MCMC)
  - PhyML (maximum likelihood)
  - RaxML (fast maximum likelihood)
  - FastTree2 (very large heuristic trees)

Finish your dynamic programming implementation.