

Multiple Alignments and Phylogenies

Mark Voorhies

3/29/2012

Comprehending our BLAST results

- We have a bunch of sequences that look similar to our query

Comprehending our BLAST results

- We have a bunch of sequences that look similar to our query
- We infer that they are homologous to each other

Comprehending our BLAST results

- We have a bunch of sequences that look similar to our query
- We infer that they are homologous to each other
- What does that mean, anyway?

Homologs heritable elements with a common evolutionary origin.

Nomenclature

Homologs heritable elements with a common evolutionary origin.

Orthologs homologs arising from speciation.

Paralogs homologs arising from duplication and divergence within a single genome.

Nomenclature

Homologs heritable elements with a common evolutionary origin.

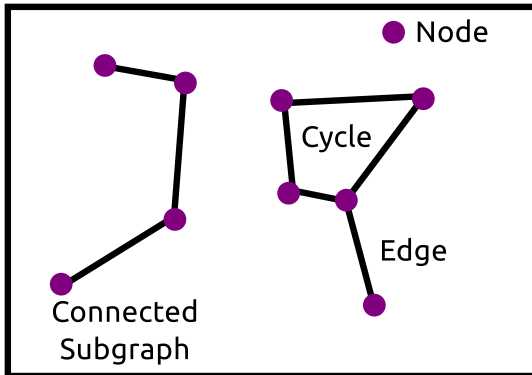
Orthologs homologs arising from speciation.

Paralogs homologs arising from duplication and divergence within a single genome.

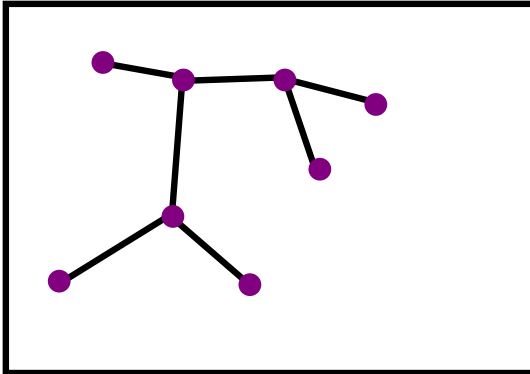
Xenologs homologs arising from horizontal transfer.

Onologs homologs arising from whole genome duplication.

Graph



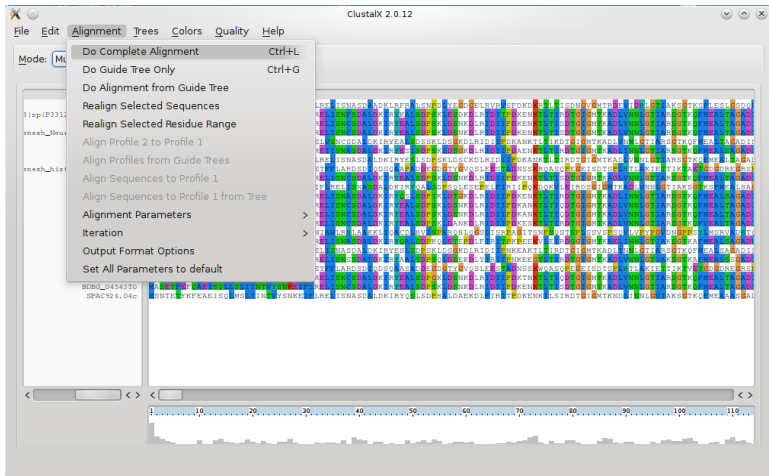
Tree = Connected Graph with no Cycles



Generating a multiple alignment in CLUSTALX

The screenshot shows the ClustalX 2.0.12 interface. The 'File' menu is open, showing options like 'Load Sequences', 'Append Sequences', 'Save Sequences as...', 'Load Profile 1', 'Load Profile 2', 'Save Profile 1 as...', 'Save Profile 2 as...', 'Write Alignment as Postscript', 'Write Profile 1 as Postscript', 'Write Profile 2 as Postscript', and 'Quit'. The main window displays a multiple sequence alignment of protein sequences. The sequences are color-coded by amino acid type. A sequence list on the left includes: A009010200062D, _CMAQ_04150_1, BCEE_031407D, PAAQ_046797D, BDBI_045437D, and SPAC92.6.04c. The alignment shows conserved regions across the sequences, with a scale bar at the bottom indicating positions from 1 to 110.

Generating a multiple alignment in CLUSTALX



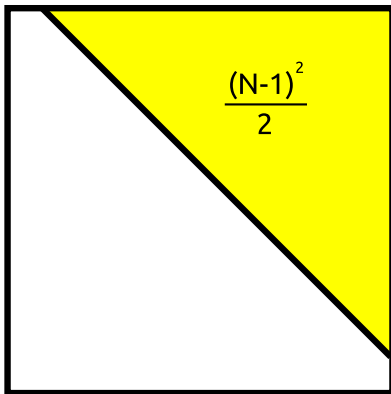
Evolution implies a self-consistent model



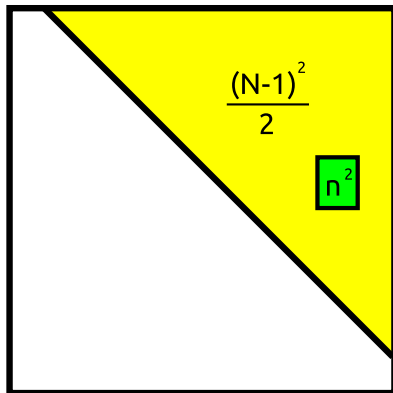
Distances
(Pairwise relationships)

Topology
(Evolutionary history)

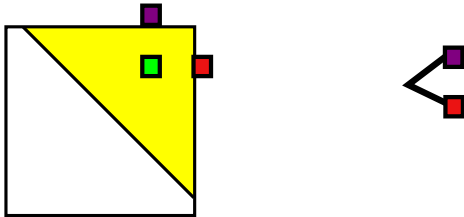
Measure all pairwise distances by dynamic programming



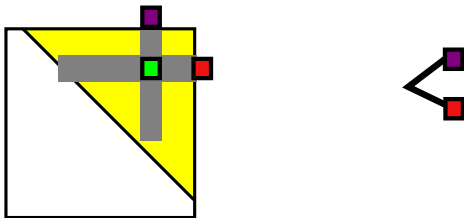
Measure all pairwise distances by dynamic programming



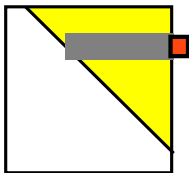
Generate a guide tree by UPGMA



Generate a guide tree by UPGMA



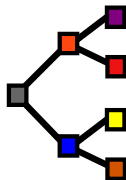
Generate a guide tree by UPGMA



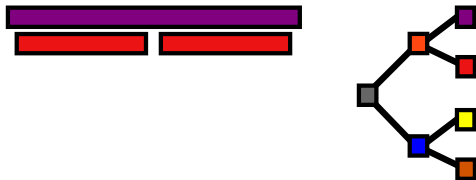
Generate a guide tree by UPGMA



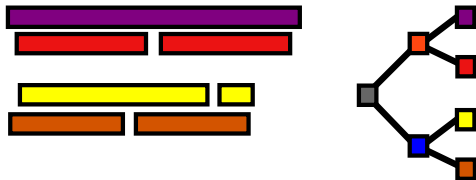
Generate a guide tree by UPGMA



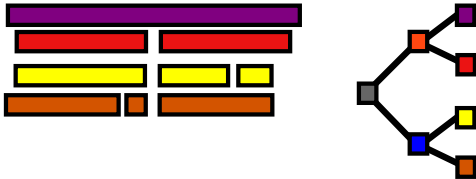
Progressive alignment following the guide tree



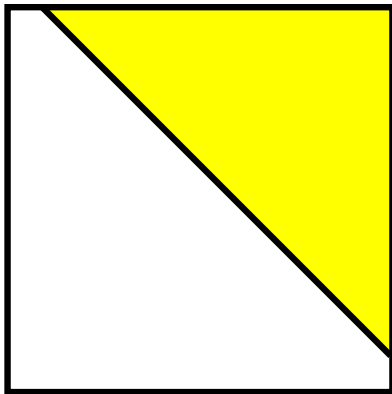
Progressive alignment following the guide tree



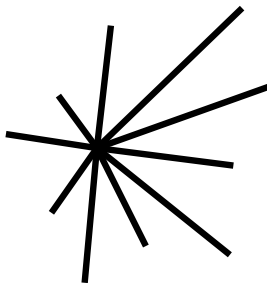
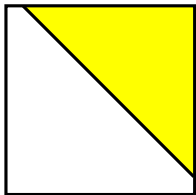
Progressive alignment following the guide tree



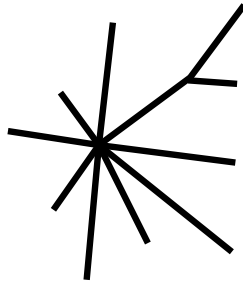
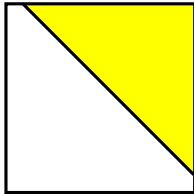
Measure distances directly from the alignment



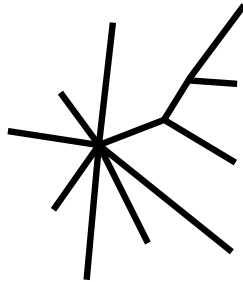
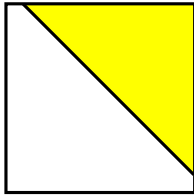
Generate neighbor-joining tree from new distances



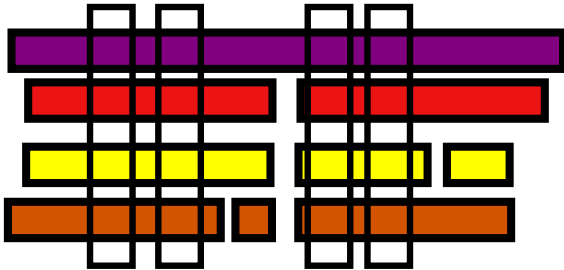
Generate neighbor-joining tree from new distances



Generate neighbor-joining tree from new distances



Generate bootstrap values from subsets of the alignment



Generating a neighbor joining tree in CLUSTALX

The screenshot displays the ClustalX 2.0.12 application window. The 'Trees' menu is open, showing options: 'Draw Tree' (Ctrl+R), 'Bootstrap N-J Tree' (Ctrl+B), 'Exclude Positions with Gaps', 'Correct for Multiple Substitutions', 'Output Format Options', and 'Clustering Algorithm'. The main window shows a multiple sequence alignment of protein sequences, with a color-coded background for each amino acid. A progress bar at the bottom indicates the alignment is complete, with the text 'CLUSTAL-Alignment file created [Hsp82aa.aln]'.

ClustalX 2.0.12

File Edit Alignment **Trees** Colors Quality Help

Mode: Multiple Alignment

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

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. The alignment is color-coded by conservation, with a 'Conservation Colour Increment (Background)' dialog box open, allowing the user to adjust the visibility of conservation (currently set to 55). The alignment includes a 'Conservation' bar at the top, a 'Quality' bar, and a 'Consensus' bar at the bottom. The sequences are listed on the left, and the alignment is shown in the center. A 3D ribbon diagram of a protein structure is visible on the right side of the interface, labeled 'Jmol'.

File Tools Help Window

Overview /home/mvoorhies/data/Sinem/Antibodies_2_11_2010/Hsp82aa.aln

/home/mvoorhies/data/Sinem/Antibodies_2_11_2010/Hsp82aa.aln

File Edit Select View Format Colour Calculate Web Service

1820 1850 1880 1910 1940 1970 2000

HCAG_046862-702 --MSS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

ACDQ_03083702-2445 --AAMS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

Contig0_30_Fgnesesh_Neurospora_1-702 --MSS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

HCIG_03340702-1221 --AAMS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

gi417153|sp|P33125.1|HSP82_A1/-679 --MSS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

INSTD_FE_Contig19_Fgnesesh_insl2-1614 --AAMS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

BCDQ_011599702-704 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

BOBQ_04548702-704 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

PADQ_07715702-671 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

PAGQ_06249702-295 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

PAGQ_05679702-495 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

CING_047292-702 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDSEKTL

Alu3g042102-706 --MSS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

HC0901020006201-699 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

AN809_2/2-700 --MAS-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

JMGQ_06759_5_2-702 --MA-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

JNC00142_2_1-705 --MA-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

gk131_65252-707 --MADAKV-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

YMR186W2-705 --MAG-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

SPAC206_0nc2-704 --MSN-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

CHAD_06150_1_1-699 --MST-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

210q_A1/-624 --MGK-- --D-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDANKTL

Conservation Colour Increment (Background)

Enter value to increase conservation visibl...

55 Apply to All Groups

Conservation

Quality

Consensus

--AAM+3-- --ETFEFAEISQLLSLIINTVYSNKEIFLRELINSGDALDKIRYEALSDPSKLDNSKDLRIDIPDKENKTL

Sequence 5 ID: gi417153|sp|P33125.1|HSP82_A1

Jmol

- Multiple Alignment
 - T-Coffee
 - MUSCLE
 - COBALT
- Tree building
 - MrBayes (Bayesian MCMC)
 - PhyML (maximum likelihood)
- Work benches
 - MESQUITE
 - UGENE

Searching with PSI-BLAST

Protein BLAST: search databases using a protein query - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE=Proteins&PR psi-blast

Protein BLAST: search databa...

Or, upload file Browse...

Job Title
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Organism
Optional
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. Exclude

Exclude Models (X/M/XP) Uncultured/environmental sample sequences
Optional

Entrez Query
Optional
Enter an Entrez query to limit search

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST Search using PSI-BLAST (Position-Specific Iterated BLAST)

Show results in a new window

[Algorithm parameters](#)

Done

- Play with CLUSTALX, JALVIEW, and PSI-BLAST
- Read PLoS Comp. Biol. 4:e1000069