Massachusetts Institute of Technology
6.877 Computational Evolutionary Biology, Fall, 2005
Laboratory 3: Detecting selection

Handed out: November 28 **Due: December 14**

Introduction

In this laboratory we shall apply evolutionary models to genomic data – from the very lowest levels, nucleotides and amino acids, to the question of whole-genome duplication. Our major goal will be to have you become familiar with modern methods for detecting natural selection at the level of genes and proteins, as well as to engage some current recent questions in gene duplication. The lab will consist of two parts: the first will have you get familiar with the tools, including volatility analysis, while the second turns to the maximum likelihood method, which is the current 'state of the art.'

Part 1: Detecting selection – two basic methods and a new, noncomparative approach

Part 1.1 Detection of selection using synonymous/nonsynonymous ratios in primate lysozyme enzymes

Objective: Analyze the pattern of nonsynonymous and synonymous substitutions between genes.

Dataset:

Already-aligned lysozyme C precursor DNA sequences from different species of apes (including human) in FASTA format may be found in the labs section. Please download and save this file to your local computer, for use in the next step.

Procedure:

You can use the perl-based SNAP utility at the HIV database:
http://hiv-web.lanl.gov/SNAP/WEBSNAP/SNAP.html.

This program will calculate overall numbers for K_A and K_S (it uses the terms dN dS instead), as well as a cumulative plot of these values, codon by codon as we move through a coding region, as well as some simple statistics for these counts. It can also use the dN and dS values as a 'distance' metric to construct a 'nearest neighbor' or 'neighbor joining' (NJ) phylogenetic tree relating the species that you are comparing.

Procedure:

Here is how to use it once you've navigated to this site (see snapshot below):

1. For "Format", select "Fasta" in the drop-down box.
2. Use "Choose file" to select the local fasta file on your computer that you've already downloaded.
3. Check <u>all</u> the next four boxes to select all the analyses and the two possible "Neighbor joining" (NJ) phylogenetic trees, one based on synonymous substitutions; one based on nonsynonymous.
4. Hit the "Run Snap" button.

The program will do its calculations and then display a page with five links to the results.

This program uses the Nei-Gojobori 'corrected path counting' method, which adjusts for counts via Jukes-Cantor plus the weighting of pathways from one codon to another according to an equi-probable model for each possible codon-to-codon path; for details on this approach, you can read the original paper Nei, Masatoshi, and Takashi Gojobori. "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." *Molecular Biology and Evolution* 3, no. 5 (1986): 418-426.

Please also take a look at theREADME link for the SNAP utility, here:
http://www.hiv.lanl.gov/content/hiv-db/SNAP/README.html.

If you prefer to do some of your own hacking with this method, the perl source code is available for Unix, Macs, and PCs, at these ftp locations:

For Unix: ftp://ftp-t10.lanl.gov/pub/aids-db/PROGS/Snap
For Macintosh: ftp://ftp-t10.lanl.gov/pub/aids-db/PROGS/Snap.mac
For PC: ftp://ftp-t10.lanl.gov/pub/aids-db/PROGS/Snap.pc

Questions.

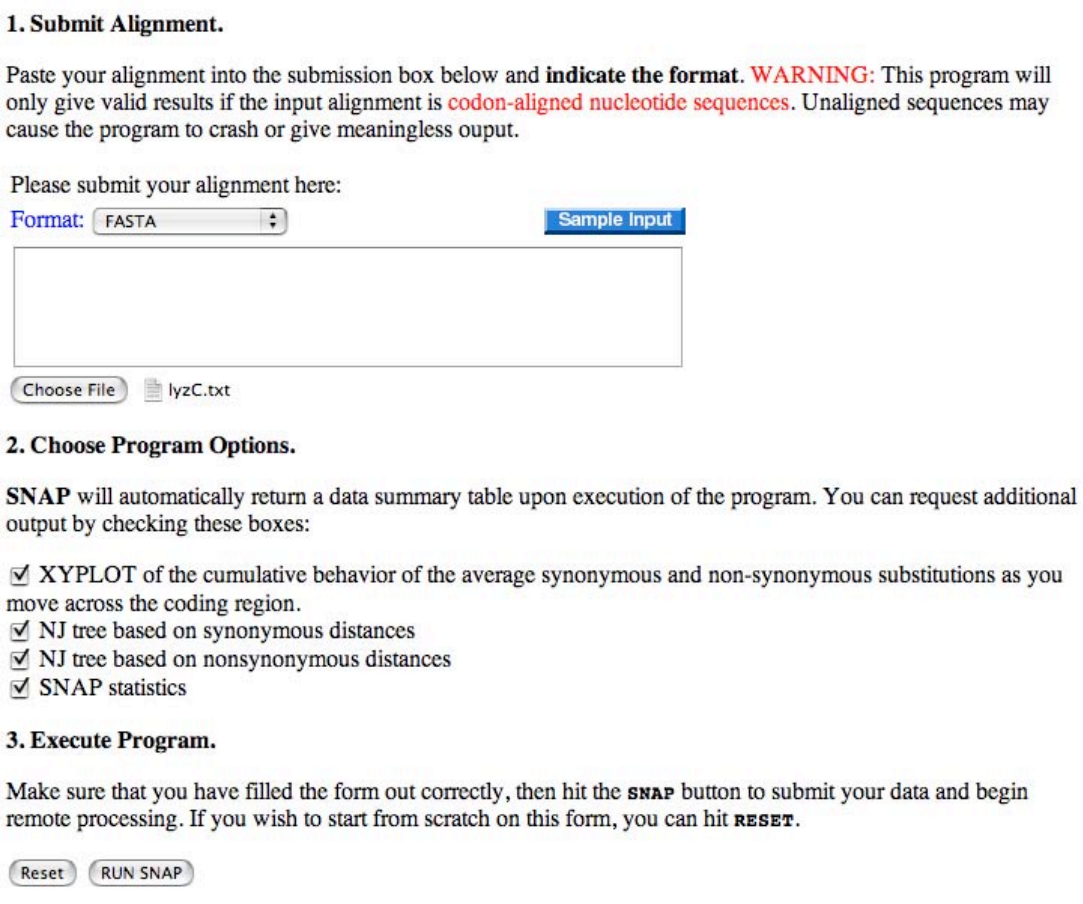
1. Using the nonsynonymous/synonymous ratio test, is there any evidence for positive selection on the lyz C gene in ape-like primates?

2. What is the overall dN/dS ratio for the lyz C genes in the alignment? Is there an excess of silent or coding mutations?

3. Are the phylogenetic trees based on dN and dS different? What do you think this means?

4. Discuss the results of this analysis. Why might the lyz C genes among these species differ? Given the results, how strong is the evidence that natural selection is responsible? (Reference: Polley SD, Conway DJ. 2001. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics*, **158**:1505-1512.)

Screen snapshot for SNAP:



Courtesy of Los Alamos National Laboratory.

Part 1.2 Detecting selection with Tajima's D test

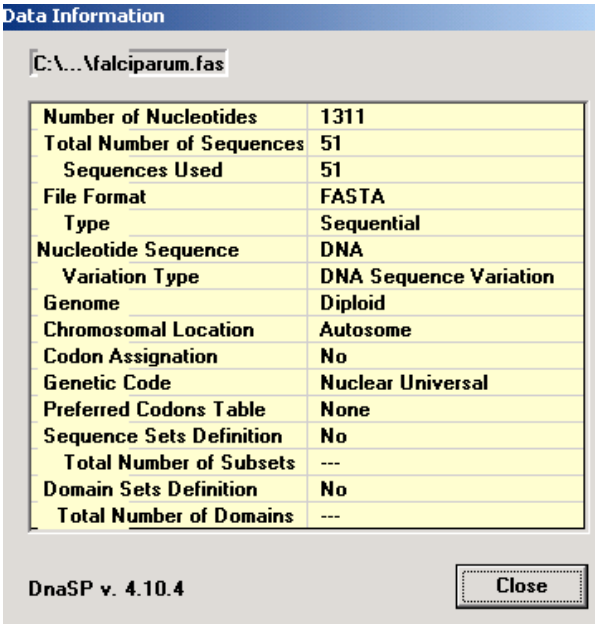
Objective:
You've already had a chance to use this polymorphism-type detector for positive selection in the previous laboratory. Now we apply it in a more systematic way, comparing methods.

Dataset:
Fetch the sequence alignment of *Plasmodium falciparum* (malaria) apical membrane antigen 1 gene sampled according to different geographic regions from the labs section and save the sequences on your local computer.

Procedure:
(For the next part we assume that you all have access to PCs running Windows…if not, please see me; it may also run on a Mac under VirtualPC, and, possibly, under Unix using WINE.) Download and install *DnaSP* on your local PC; the web site is:
http://www.ub.es/dnasp/dnasp4104.exe.
The main web site is: http://www.ub.es/dnasp/.
This file is a self-extracting, and you just run it to install the program itself. You will use the program to analyze nucleotide diversity and test selection with respect to this dataset via Tajima's *D* but now using a sliding window of appropriate size.

You can then use DnaSP as follows.

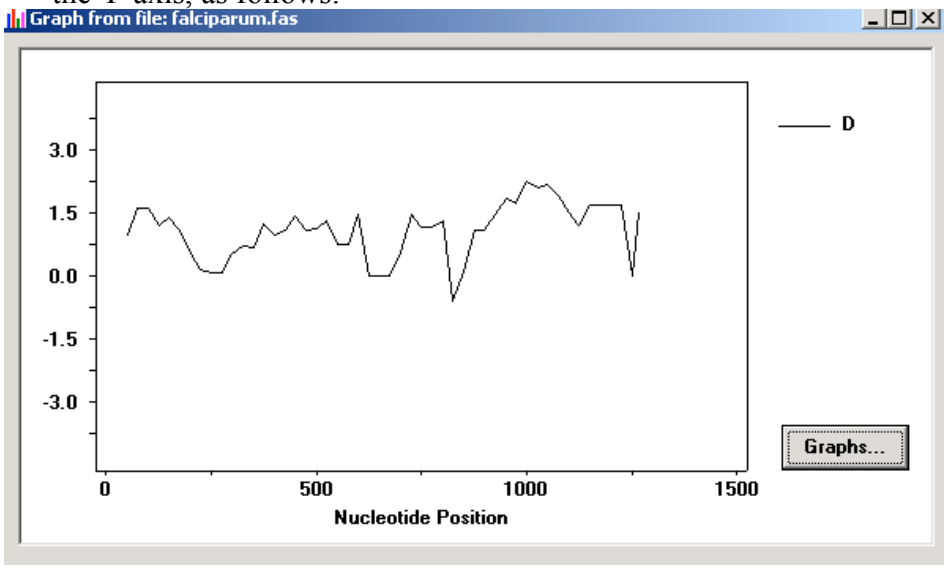
- Start the program DnaSP and from the "File" menu open the `faliciparum.fas` file that you've saved locally. The program will put up a short summary window, which you can then close. The window looks like this:



Courtesy of Julio Rozas. Used with permission.

- Select "Overview" from the menu panel, selecting the only item in the pull-down menu, and click "OK" to the resulting dialog box (which selects all 1311 sites to compare) and run the program to obtain summary statistics that include an estimated number of 'haplotypes' as well as summary calculations of selection neutrality (Tajima's D, Fu and Li's D*, etc.) – record these. Note that these are calculations for the gene *as a whole*.

- Carry out a sliding window analysis of the nucleotide diversity: under the "Analysis" tab select "Tajima's Test" and then from resulting pop-up Options window, look on the right where one selects a 'sliding window' analysis, so that one can see how *D* varies from gene region to gene region. Check the 'Compute' box – but you will have to do some experimenting to figure out the step size should be (try the default first?). After you click "OK" you will get a pane that has text output and a table. The table will have a check-box to produce a graph that plots nucleotide (DNA) position on the X-axis and Tajima's *D* on the Y-axis, as follows:



Courtesy of Julio Rozas. Used with permission.

Questions.

1. What does Tajima's D tell you about this gene as a whole? What form of selection is this? (Neutral, diversifying, purifying, balancing?) Please briefly explain. What do any of the other Tajima-type tests say differently, if anything? (These include Fu and Li's D*, Fu and Li's F* – we haven't covered these in detail, but, briefly, Fu and Li's Tests DnaSP computes the D, D*, F and F* test statistics proposed by Fu and Li (1993) to test various predictions made by the neutral theory of molecular evolution (Kimura 1983). The test statistics D and F require data from intraspecific polymorphism and from an outgroup (a sequence from a related species), and D* and F* only require intraspecific data. DnaSP uses the critical values obtained by Fu and Li (1993) to determine the statistical significance of D, F, D* and F* test statistics. DnaSP can also conduct the F test statistic (Fu 1997).

2. Using the sliding window method, see if you can locate any region in the gene that seems to be under selection, as opposed to the gene as a whole (below we ask you to see if you can figure out why these sites might be under selection).

Here is some detailed information on this gene, *P. falciparum* PF11_0344, an apical membrane precursor, 622 amino acids long. It resides on Chromosome 11 of *P. falciparium*, spanning locations 1290767 – 1292635. I have placed a link to the NCBI entry on this gene and a link to the plasmodium database entry for this gene below. You can use these links to view the gene's amino acid sequence, nucleotide sequence, and even the 3-D structure of part of the corresponding protein (structure 1HN6). Some snapshots from the plasmodium database are below. Perhaps the best description is the link that follows, from the TIGR, which gives a good overview of the functioning of the protein associated with this gene and a great browser view.

Link 0: TIGR Center *Plasmodium falciparum* database
http://www.genedb.org/genedb/Search?submit=Search+for&name=PF11_0344&organism=malaria&desc=yes&wildcard=yes

Link 1 to the NCBI genome database:
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&val=23508535

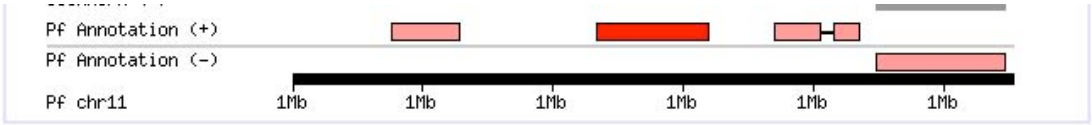
Link 2 to the Plasmodium database:
http://plasmodb.org/plasmodb/servlet/sv?page=gene&source_id=PF11_0344

Link 3 to the Plasodium database genome browser:
http://plasmodb.org/cgi-bin/plasmodb/servlet/genomicSeqBrowser.pl?geneSourceId=PF11_0344&taxonId=211

Link 4 to the 3-D crystal structure of this protein in the protein data base – PDB database:
http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics&pdbId=1HN6

Link 5 to the Conserved Domain Database (CDD) at NCBI:
http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=2940

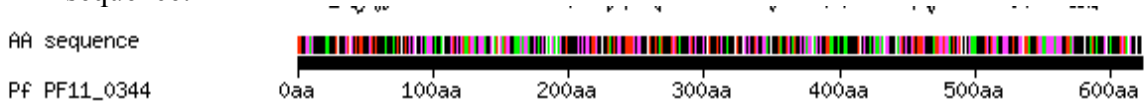
Chromosome 11 picture:



Exon locations / gene model back to top

exon	location	strand	length	algorithm	coding start	coding end
1	1290767 - 1292635	+	1869	Pf Annotation	1	1869

Gene AA sequence:



Courtesy of U.S. NIH and U.S. NIAID. Source: PlasmoDB (http://plasmodb.org)

Image removed due to copyright restrictions.

The amino acid sequence for this gene is as follows (from NCBI):

```
  1 mrklycvlll safeftymin fgrgqnyweh pyqnsdvyrp inehrehpke yeyplhqeht
 61 yqqedsgede ntlqhaypid hegaepapqe qnlfssieiv ersnymgnpw teymakydie
121 evhgsgirvd lgedaevagt qyrlpsgkcp vfgkgiiien snttfltpva tgnqylkdgg
181 fafppteplm spmtldemrh fykdnkyvkn ldeltlcsrh agnmipdndk nsnykypavy
241 ddkdkkchil yiaaqenngp rycnkdeskr nsmfcfrpak disfqnytyl sknvvdnwek
301 vcprknlqna kfglwvdgnc ediphvnefp aidlfecnkl vfelsasdqp kqyeqhltdy
361 ekikegfknk nasmiksafl ptgafkadry kshgkgynwg nyntetqkce ifnvkptcli
```

```
421 nnssyiatta lshpievenn fpcslykdei mkeiereskr iklndnddeg nkkiiaprif
481 isddkdslkc pcdpemvsns tcrffvckcv erraevtsnn evvvkeeykd eyadipehkp
541 tydkmkiiia ssaavavlat ilmvylykrk gnaekydkmd epqdygksns rndemldpea
601 sfwgeekras httpvlmekp yy
```

The full open reading frame (ORF) sequence (so far) of the entire *falciparum* chromosome 11 may be found here in fasta format. Please download it from the labs section to your local machine, for the next part of the lab.
(Note: this is a 1.5 MB download!)
(This includes just predicted ORFs larger than 50 amino acids long.)

The full nucleotide sequence for gene AMA1 (PF11_0344) may be found in the labs section.

Part 1.3. Codon volatility and selection

Objective:
We will next attempt to test in a noncomparative way whether this apical membrane gene is under positive selection.

Dataset:
As in the previous section.

Procedure:
You will use J. Plotkin's volatility server, at http://volatility.cgr.harvard.edu/cgi-bin/volatility.pl, and run it over all of the ORFs in chromosome 11 to see whether the PF11_0344 gene stands out. (The server takes fasta format as input, so you can use the chromosome 11 sequence in the lab directory, above.) Please go ahead and do this now, as follows: Go to the volatility server and upload the entire chromosome 11sequence from your local computer that you had downloaded above. Next…

Questions.

1. What do you think the kappa value (transition/transversion rate ratio) should be? Take a look at the kappa values used for other genomes that have been already placed on this server, by selecting the "Upload a fasta file" pull-down menu – you'll note that th for the whole *falciparum* genome Plotkin used a kappa value of 1.0 (Click on the 'About kappa' box and you will see that for the example sequences on this site, kappa has been estimated via a maximum likelihood analysis. See if you can figure out how to use DnaSP to estimate kappa, or do a bit of research on how this is done.)

2. Run the volatility analysis. After you run the program, you can download the result to an excel spreadsheet locally. Then you'll have to do a bit of grepping to find the PF11_0344 gene (but it's not that hard). See what its computed volatility is – does it agree with your previous analysis from section 1.2? How does demographic change enter into this picture (if at all)? What do you think this agreement or disagreement means? The citation for Plotkin's Nature paper describing his method is Plotkin, J. B., J. Dushoff, and H. B. Fraser. "Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum." *Nature* 428, no. 6986 (April 29, 2004): 942-945.

Screen snapshots (first a shot of the input page, then the top of the sample result page). Note that the tabulation from the web output is a bit 'off', but it should be downloaded into a spreadsheet anyway, as suggested on the web page itself.

Welcome to the Codon Volatility Computation Server (v1.0)

This public web server allows users to calculate the volatility and associated volatility P-values for nucleotide coding sequences. The volatility P-values of genes reflect the relative selective pressures for or against amino acid changes, as described in the accompanying paper:

J. B. Plotkin, J. Dushoff, and H. B. Fraser.
Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum.
Nature, 428: 942-945 (2004).

Enter a name for this analysis (optional): test

Please select a genome whose volatility you wish to compute. Upload a FASTA file.

Or upload a file in FASTA format: Choose File PFA1_chromo_11.fasta About Input Files

Specify the transition-transversion bias, kappa: 1.0 About Kappa

RUN VOLATILITY ANALYSIS Reset About Output

More information about the volatility computation can be found here:

Analysis Results for Volatility

Right click here to save output file as text.

```
Output from Volatility 1.0 (c) J.B. Plotkin:

 Parameter settings:
 #cdfile        /usr/local/www/prod/public/volatility/volatilityweb/output/test_17525/test_17525.cd (command)
 #TagLength     5000
 #MaxGenes      40000
 #trials        1
 #trialreps     0
 #Report        100
 #MinHits       5
 #seed  257
 #KAPPA 1.000000 (command)


 Read 1009 genes from FASTA file.

 Name   Length  Observed        Expected        Variance        P-value
 >539.t00001 hypothetical protein       1013    0.8220710133    0.8229936641    0.0014278878    0.7814596750
 >540.t00001 MAL3P7.34            (PFC1015c), Hypothetical protein, len: 2341 aa       2481    0.8218836171
```

Courtesy of J. Plotkin. Used with permission.

Part 1.4. Going further

Let's do a bit more digging on this one – a more open-ended question. The analysis in section 1.2 looked at parts of a single gene, while the volatility analysis compares whole genes against the rest of a genome. If you have located the regions within this gene that seem to be undergoing selection, look at which amino acid residues they correspond to. Can you locate them in the actual 3-D structural map that the protein structure database PDB provides? (I know you might have to ask your resident bioinformatics person in class about some of the details about how to do this, or the biochemistry – that is OK – just do your best and please work together.)