LAB 5: SELECTION ANALYSIS

[SOFTWARE NEEDED: web access]

Last lab, we took an alignment of homologous bacterial genes and created phylogenetic trees, which allowed us to make inferences about how they are related and how they may have evolved. Now we want to look at the impact of natural selection on these genes. In this lab we will attempt to determine if selection is acting, what kind of selection, and how the sequences changed because of it.

Why do we care about the action of natural selection? Is it important to understand if you are strictly interested in studying gene function? Natural selection is the fundamental mechanism underlying all adaptive change. The effective and efficient functioning of genes, proteins, and their interactors is largely due to the action of natural selection. Therefore, we can gain great insight not only into how genes and proteins evolve by the study of natural selection, but can also identify the genetic changes responsible for specific adaptations by identifying the patterns left by natural selection on the genome.

Natural selection is very simply the process by which heritable genetic variants change in frequency due to their impact on the fitness of the organisms carrying them. The key with natural selection is *variation* – natural selection can only act when there is heritable genetic variation to act upon. Some of this variation may be 'good', and therefore may increase in frequency, while other variation may be 'bad', and may therefore decrease in frequency. 'Good' and 'bad' should not be over-interpreted, but for example may simply mean that a specific variant of a protein functions more effectively in a particular environment. That same variant may in fact function less effectively in a different environment, so natural selection is not only dependent upon the presence of heritable genetic variation, but also the particular environment where that variation is found. Natural selection should also *not* be thought of as some sort of acting entity. It is, very simply, the differential survival and transmission of genetic variants to following generations.

Box 1. The Major Types of Natural Selection

Natural selection can act in a wide range of different manners. The simplest and most relevant for this course are listed below:

- Positive selection occurs when a beneficial mutation occurs in a population and increases in frequency. Of course, if this mutation is increasing in frequency, then other variation must be concurrently decreasing in frequency. The classic example of positive selection is the spread of an antibiotic resistance allele through a population of bacteria that are being exposed to that antibiotic.

- Negative (purifying) selection is the opposite of positive selection. It occurs when a detrimental mutation is selected out of a population. Since most proteins have undergone millions or billions of years of evolution, most mutations that cause a change

to a protein coding sequence are believed to be detrimental. These detrimental mutations will be purged from the population by negative selection. For any protein, the fraction of mutations that are detrimental is directly related to the 'evolutionary conservation' of that protein.

- Balancing or diversifying selection is selection that favours the maintenance of genetic variation at a locus. While both positive and negative selection purge variation (either selecting *for* or *against* variants), balancing selection actually maintains variation by selecting for multiple genetic variants. The easiest way to imagine this happening is to consider the case of multiple environments that select for different allelic forms of a protein. For example, a receptor protein that is strongly beneficial when a pathogen is present in a population, but does nothing but put a load on the system when there is no pathogen.

How do we detect and measure natural selection? Fortunately for us, all evolutionary processes (including natural selection, mutation, recombination, gene flow, and genetic drift) leave their characteristic marks or footprints on the genome. Some of these footprints are obvious and long-lasting, while others are obscure and / or very transient. If you know what to look for, and where to look then you can very often reconstruct the evolutionary history of a genetic region.

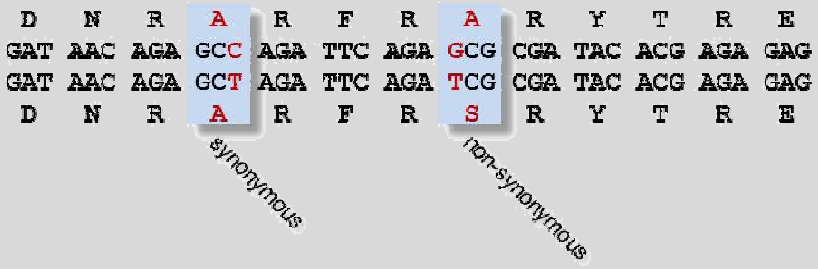
There are many approaches for identifying and characterizing the footprints left behind by natural selection. For our purposes, we are going to focus on the most common test to identify positive and negative selection. As mentioned, positive selection is selection for a genetic change that increases fitness (benefits the organism), while negative selection is typically associated with evolutionary conservation to preserve an essential function.

dN/dS Ratio Test

The dN/dS Ratio Test is perhaps the most widely used method for detecting the pattern of natural selection from nucleotide sequence data. This test is particularly useful because it can infer selection acting all the way down to the level of the codon.

Box 2. Using dN and dS to Infer Selection

The dN/dS (also known at the Ka/Ks or ω) test calculates the ratio of the rate of non-synonymous substitutions (dN, the number of non-synonymous substitutions per non-synonymous site) to the rate of synonymous substitutions (dS, the number of synonymous substitutions per synonymous site). Non-synonymous substitutions are those mutations that result in a change in the protein sequence, while synonymous substitutions are those that change the DNA sequence, but not the protein sequence due to the degeneracy of the genetic code. Note that we are interested in the *rate* of these substitutions, not their absolute number.



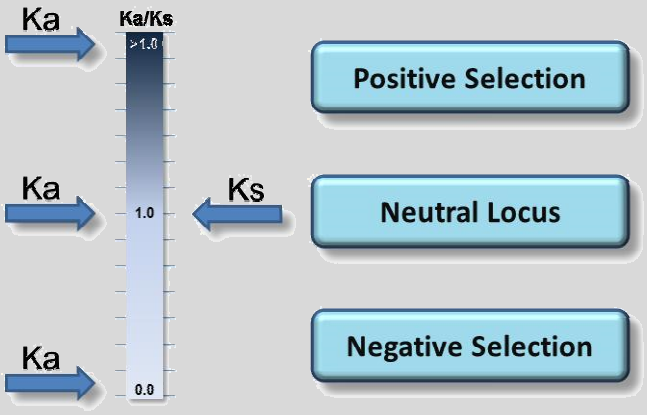
Synonymous substitutions are generally not exposed to strong selective pressures since they don't result in a change to the protein sequence (although there is some evidence that some codons are preferred over others for efficient translation to occur); therefore, they tend to accumulate at roughly a constant rate. Think of this rate as the baseline by which we will compare the rate of substitutions that change the protein sequence (non-synonymous substitutions).

In the case of a completely neutral sequence (one that is free to change with no constraints), you would expect dN to be the same as dS, or the ratio dN/dS = 1.

When there are selective constraints on a sequence (negative selection), you would expect fewer substitutions that change the protein sequence, or a lower dN; therefore, dN/dS < 1.

In the case of positive selection, you would expect to see a higher proportion of amino acid substitutions in your population (because they are being increased by positive selection), so a higher dN; therefore, dN/dS > 1.

We can determine if a gene is under positive or negative selection by measuring the dN/dS ratio. dN/dS > 1 is a strong indicator of positive selection. dN/dS < 1 is a strong indicator of negative selection. A neutral sequence should, in theory, have a dN/dS = 1. The easiest way to think about this is that the rate of synonymous (silent) substitutions should remain constant since it is not exposed to selection, but the rate of non-synonymous substations can go up or down depending on whether there is selection to change the amino acid sequence or keep it the same, respectively.



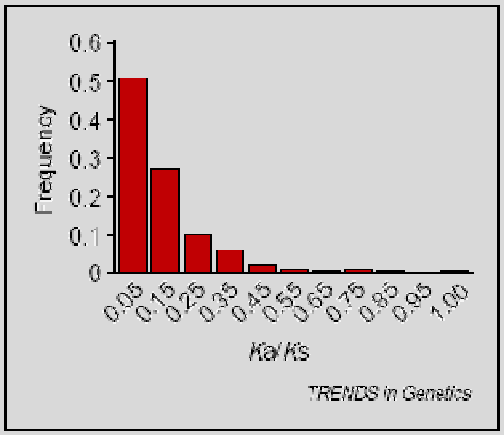


Fig. 1. The frequency of different values of *Ka/Ks* for 335 mouse–rat orthologous genes. Figures on the *x* axis represent the middle figure of each bin; that is, the 0.05 bin collects data from 0 to 0.1

Most functional genes are under some level of selection constraint; therefore, dN/dS ratios are typically well below 1.0 for most coding regions, see Figure 1 from Hurst (2002) TIG 18:466 to the right.

We can also use the dN/dS test to identify selection acting at the level of individual amino acids. This can be useful for identify functionally important regions – you could imagine that if there's an "arms race" between a host and a pathogen mediated by a host protein and pathogen protein interaction, that the surface where the two interact could be under positive selection whereby the rest of the protein could be under negative selection to maintain the overall biochemical function of the protein. For such tests to work it is necessary to have an appropriate evolutionary likelihood model, which could allow certain sites within the sequence to be under positive selection, with the majority of the sites being under negative selection. Different types of models are used in Datamonkey and other selection analysis tools, such as PAML.

We will be working with a set of aligned sequences for the HrpZ gene from several strains of *Pseudomonas syringae*, a plant pathogen. HrpZ encodes a component of the type III secretion system of this bacteria, which is important for delivering proteins that can disrupt the plant's immune system. Download **bioinfomethods1%2Flabs%2FLab5_Psy_hrpZ.fas** from the Coursera website (get it from the part of the Coursera site where you downloaded this lab manual; right- or Command-click on the "here" link there to download the alignment in Fasta-format to your computer for use in this lab. If you're a Mac user, note that a .txt extension will be added to the filename).

Obviously, it is extremely important to work with DNA sequences that have been aligned according to their amino acid alignments. *Why*? See the first figure in Box 2 if this is not clear. It is often useful to check your aligned DNA sequences using the TRANSEQ app of EMBOSS suite of tools from the EBI. Go to http://bar.utoronto.ca/EMBOSS (a local instance) then sort the functions alphabetically using the link in the top left to find the TRANSEQ function. Paste the sequences in the Lab5_Psy_hrpZ.fas file into the box, and choose the 1st reading frame to confirm that **all sequences are in frame and contain no stop codons**.

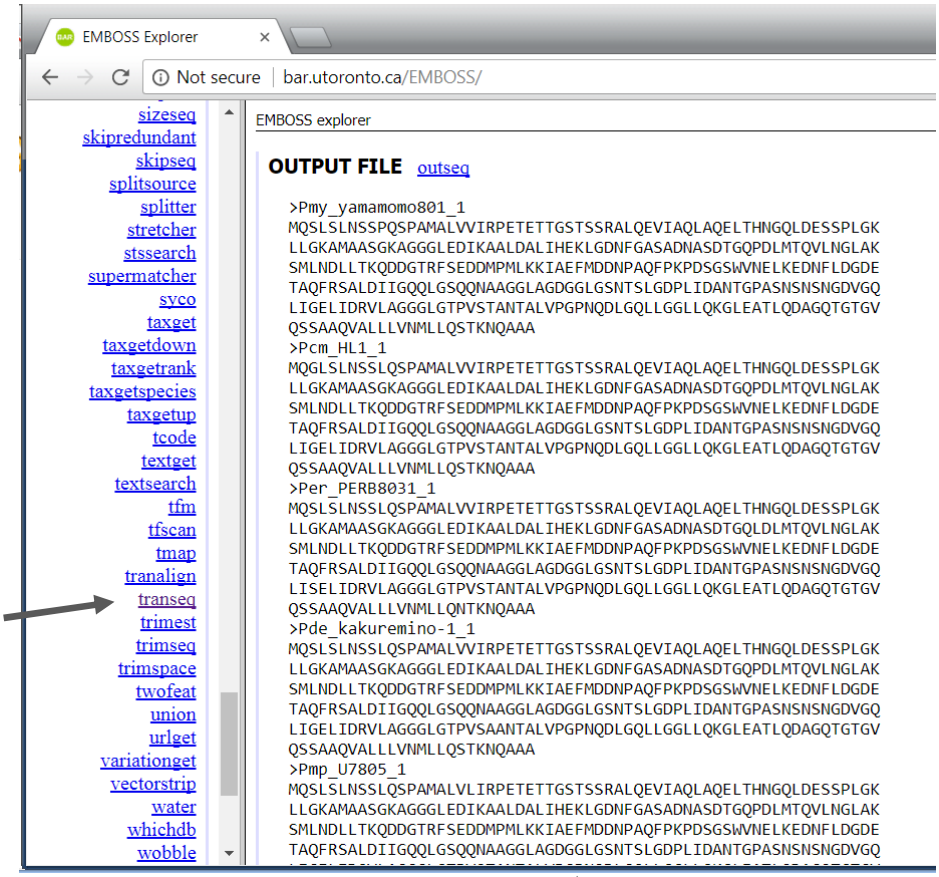


Figure 1. EMBOSS "transeq" output for 1st reading frame of aligned HrpZ file. Note that all sequences start with a methionine, and that there are no stop codons anywhere.

We will use the online tool **Datamonkey** (http://www.datamonkey.org/) to look for selection in our sequences. Datamonkey is a very straightforward and powerful tool that provides access to complex and sophisticated evolutionary analyses. These analyses are carried out on a remote server so you don't need access to a high-powered workstation. Datamonkey uses the HyPhy package (Pond *et al*., 2005), which is a very powerful multiplatform package for carrying out likelihood-based analyses of rates and patterns of sequence evolution.

Extensive information about Datamonkey can be found at their website, along with a very complete help page (http://www.datamonkey.org/help). Note that we have borrowed very heavily from the original Datamonkey tutorial for this lab.

We will analyze 45 sequences of the *Pseudomonas syringae HrpZ* gene. Datamonkey provides a range of different analyses which can be selected using the **Method** drop-down menu. We will focus on the two most straightforward analysis for this lab, SLAC and FEL (if you start with the Datamonkey homepage, the decision tree answers to get to the FEL analysis are Evolutionary process: *Selection*; Detect selection at: *Sites*; Detect: *Pervasive selection*; Dataset: *Small* – you can see that we are just skimming the surface of Datamonkey's capabilities in this lab), although we encourage you to look through the references/citations provided on the site for information on more sophisticated analyses. The SLAC (single-likelihood ancestor counting) and FEL (fixed effect likelihood) methods were shown to

perform approximately the same, see the Pond and Frost (2005) reference, whereby the authors recommend SLAC for larger data sets (>40 sequences) and FEL for smaller data sets (20-40 sequences). SLAC is computationally more efficient, but has slightly less statistical power.

1. Make sure you have access to the data file: **bioinfomethods1%2Flabs%2FLab5_Psy_hrpZ.fas** (see previous page for download instructions).

2. Go to the Datamonkey analysis site for SLAC http://datamonkey.org/slac.

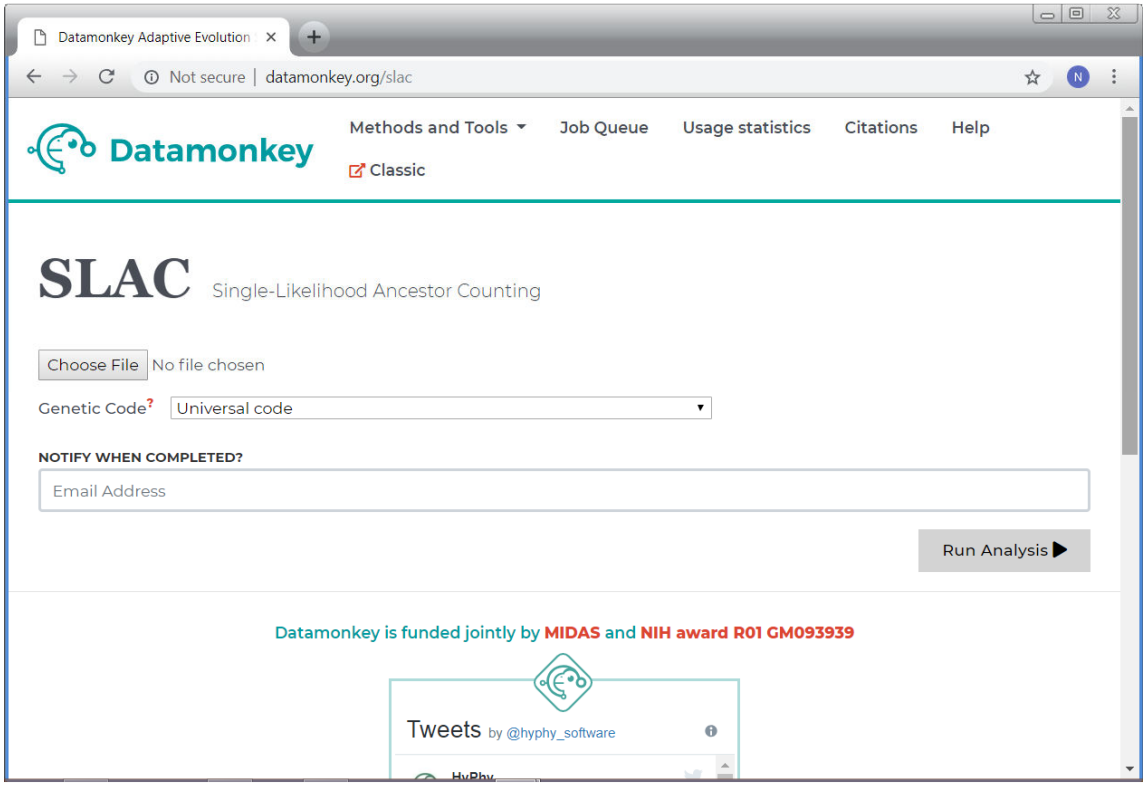


Figure 2. Datamonkey SLAC analysis upload page.

3. Upload the multiple sequence alignment file
   - Use the **Browse** button to locate the data file on your computer
   - Choose the genetic code, which would be **Universal** in this case.
   - Enter your **Email Address** and tick the box to be notified when the analysis job is complete

4. Click the Arrow button Run Analysis ▶ to proceed.

5. SLAC Results

   - After you submit the job Datamonkey will display a page showing that your job has been submitted. You'll see a job log displayed as your analysis progresses. Your

"ticket"/job number is in the URL of the page (e.g. http://datamonkey.org/slac/5a34119c0d628b50d1321433), which you can bookmark to return to a limited time later (this URL will also be sent to you once your job is done, if you've entered your email address). Then you'll see the following:

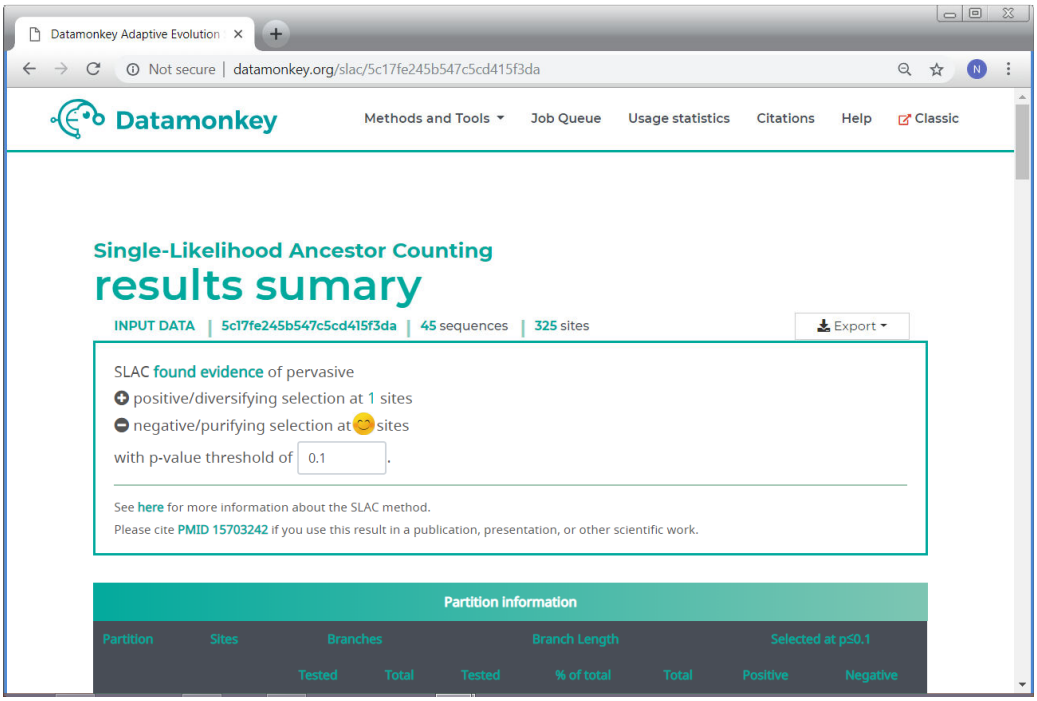


Figure 4. SLAC analysis results page

- Along the side of the results page (**Figure 4**) you will see links to more detailed results, including:
  - **Table** (can be downloaded as a comma separated values file) provides details of the analysis for each codon in the sequence. The CSV output can be easily imported into a spreadsheet program such as Microsoft Excel – we'll do this at the end of the lab, so make sure you are able to find the SLAC results page again!
  - **Graph** will bring you to a page where you can see graphical output of the analysis. For SLAC you can plot dN-dS across sites. Note that Datamonkey does not provide the more typical dN/dS plot, but plots the difference between these two instead. This is done because dS can be 0 for some sites, resulting in an infinite ratio.
  - **Tree** provides a tree of the alignment data adjusted for the substitution model used (the SLAC algorithm does a check on your data and provides the best option/s for the data).
- An important piece of data is provided in the section labeled **Partition information**, which provides the likelihood (log transformed – the lower the number the better) that the data fit the model(s) tested.

  a. *What is the likelihood for the Nucleotide GTR (general time reversible) model?*
  b. *What is the likelihood for the Global MG94xREV model?*

- In the next sections of this page, Datamonkey presents all positively (dN/dS>1) and

negatively (dN/dS <1) selected codons. Remember that dN is the rate of non-synonymous (amino acid changing) substitutions per possible non-synonymous site, while dS is the rate of synonymous substitutions per possible synonymous site; therefore, a significantly positive dN-dS value indicates positive selection, while a significantly negative value indicates negative selection. This section can be regenerated on the fly for a different significance level by entering the new level in the top section (this will also update the number of sites under positive and negative selection in the top section). Datamonkey reports the estimated dN-dS scaled by the total length of the tree (to facilitate direct comparison between different data sets), p-value for the test dN≠dS at that codon, etc.

Lab Quiz Question 1

Lab Quiz Question 2

c. How many amino acids are under positive and negative selection with the default 0.1 significance level?
d. How do the results change if you change the significance level to 0.5?
e. What amino acid position is under the strongest positive selection?
f. What amino acid position is under the strongest negative selection?

- Click **Graph** from the left panel on the page, or scroll down to that section.

- The resulting plot presents the difference between dN and dS (y-axis) for each codon along the sequence (x-axis).

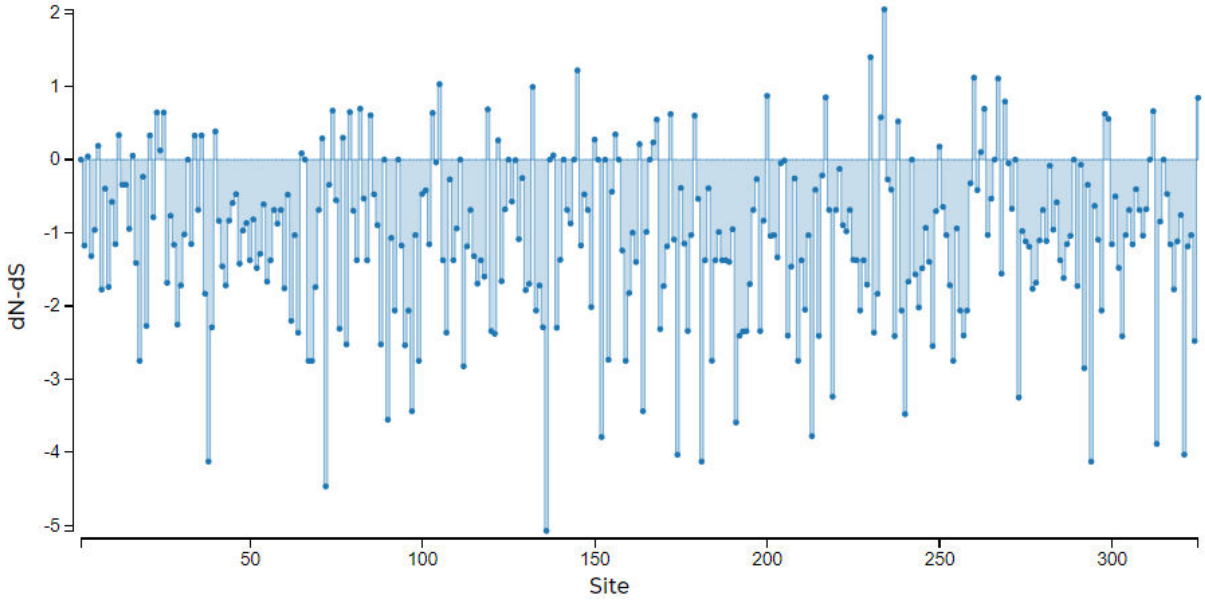


Figure 6. dN-dS plot

g. What does it mean when a codon has a dN-dS > 0?
h. What does it mean when a codon has a dN-dS < 0?
i. What does it mean when a codon has a dN-dS = 0?
j. Discuss the correspondence between the SLAC analysis of selected sites and this plot.

6. FEL Analysis
   - FEL (fixed effects likelihood) is a bit more powerful than SLAC, but is slower since it is an order of magnitude more computationally expensive.
   - Go to the FEL analysis under the Methods and Tools tab along the top.
   - Run the FEL analysis with default options using the same **Lab5_Psy_hrpZ.fas** file (Universal genetic code, then click the arrow button Run Analysis ▶ to proceed…select "All" nodes for the analysis in the intermediate tree page that appears followed by "Save Branch Selection" to initiate the analysis).

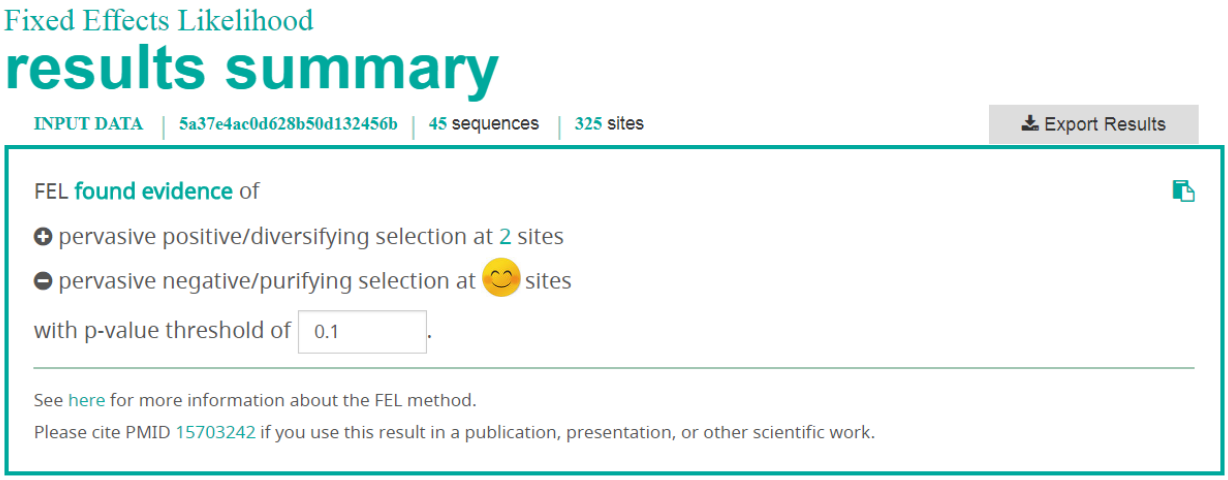


Figure 7. FEL Analysis Results

- Note the number of sites and level of significance.

Lab Quiz Question 3

m. What difference do you observe between the FEL and the SLAC analyses?

7. dN/dS Plot (optional, if you have Excel installed on your computer)

   - Now let's make a more standard dN/dS plot by exporting the Datamonkey data and performing some simple analysis in MS Excel.
   - From the **Table** section of the SLAC Analysis Results select **[Export Table to CSV]**
   - Save this page to a local file as data.csv, or similar (keep the .csv extension, though)
   - Open MS Excel
   - Drag your data.csv file into Excel – it should open automatically
   - Now you can see that this file has the following columns, among others: dS, dN, and dN-dS.
   - Sort the data by dS by selecting all of the columns (ctrl+A): In the **Data** tab select **Sort…**. In the dialogue box that pops up make sure **My List has ⊙ Header Row** is selected, and to select the **dS** column in **Sort By**. Sort by **Smallest to Largest** (i.e. by increasing values).
   - You will see about 40 rows at the top with extremely small dS values (e.g. 0 or 5.00E-09). These extremely small values will result in nonsensical dN/dS values (which is the

reason Datamonkey prefers using dN-dS). To avoid this simply delete the dN/dS values for any row with these extremely low dS values.

- Also delete the 6 rows at the bottom of the list where dS is equal to "**null**".
- Now resort the data by codon position ("**Site**").
- Compute a new column for **dN/dS** (divide the dN column values by the dS column values, across all sites) and plot it by inserting a column chart: Mark the two sets of values (**Site** and **dN/dS**) you wish to plot by highlighting the first set, then holding the Ctrl key and while doing so highlighting the second one. Then do Insert => Chart, select the Column type, and subtype Clustered column (top left). In the Wizard that pops up, remove the Series that contains the position information (Series 1) from the data Series and add it to the X-axis labels. Continue in the wizard. You should get a chart that looks like Figure 8.

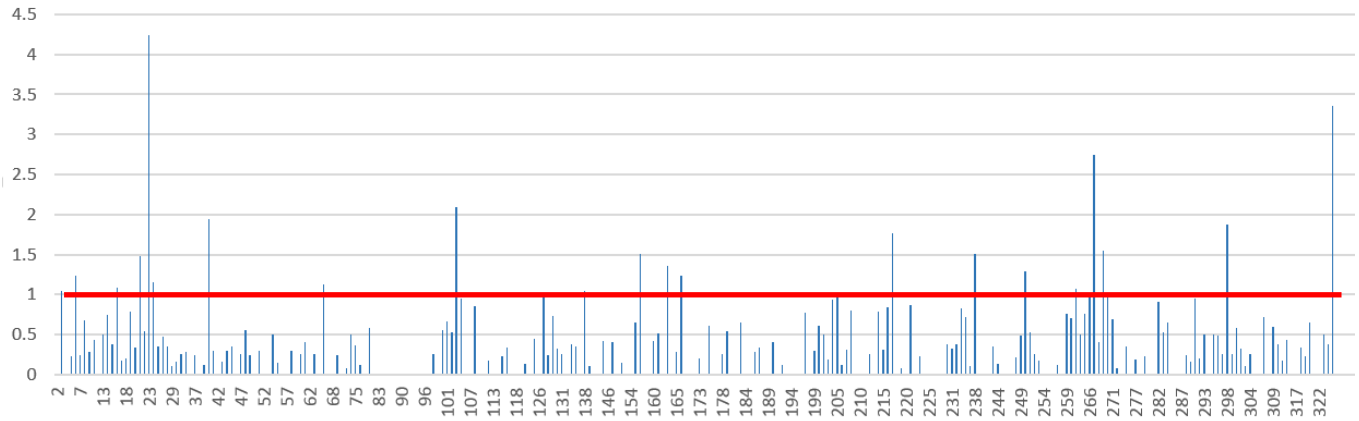


Figure 8. dN/dS plot (with dN/dS = 1 manually highlighted)

- Recall that positively selected residues have dN/dS > 1, while negatively selected residues have dN/dS<1.

n. *Are there any major differences between this plot and the dN-dS plot?*
o. *Compare the positively selected sites here to those indicated as significant positively selected sites in the SLAC and FEL analyses. Are all residues with dN/dS>1 statistically significant? Explain.*

Knowing which sites are under positive selection may be useful in e.g. designing drugs or vaccines which can act at that site in the case of a pathogen, as it tries to evade the host's defences.

- As a final exercise, let's calculate the average dN/dS score for the SLAC analysis for the sites we included in the plot in Figure 8. Use the "**=average( )**" function of Excel to calculate this.

p. *What is the average dN/dS score? Is this consistent with what you learned from the lecture?*

Lab 5 Objectives

By the end of Lab 5 (comprising the lab including its boxes, and the lecture), you should:

- know why one would be interested in ascertaining the type of selection that has been acting on a protein coding sequence, and know the three major categories of selection possible;
- understand the difference between theta (θ) and pi (π) in assessing nucleotide diversity, how these can be used to calculate Tajima's *D*, and what a given Tajima's *D* score represents in terms of selection;
- be able to calculate dN/dS;
- be familiar with how to interpret a dN/dS score;
- practically, be able to perform a selection analysis using DataMonkey, both for complete gene sequences and down to the level of individual codons.

Do not hesitate to check with the Coursera discussion forums if you do not understand any of the above after reading the relevant material.

Further Reading

Box 7.2 "The Influence of Selective Pressure on the Observed Frequency of Synonymous and Nonsynonymous Mutations" in Chapter 7 "Recovering Evolutionary History" in *Understanding Bioinformatics* by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 240-241.

Misawa K, Tajima F (1997) Estimation of the Amount of DNA Polymorphism When the Neutral Mutation Rate Varies Among Sites. Genetics 1997 147: 1959-1964.

Pond SL, Frost SD, and Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**, 676-679. www.hyphy.org

Pond SL and Frost SD (2005) Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Molecular Biology and Evolution **22**, 1208-1222.

Appendix 1: Datamonkey Usage Notes

To perform a selection analysis, Datamonkey requires an uploaded alignment of at least three homologous coding nucleotide sequences. Codon-based methods for estimating dN and dS can be applied to any sequence alignment, but there are several considerations to keep in mind.

Ideally, the alignment should represent a single gene or a part thereof (e.g. a subunit), sampled over multiple taxa or a diverse population sample. The number of sequences in the alignment is important: too few sequences will contain too little information for meaningful inference, while too many may take too long to run. At the time of this writing, Datamonkey permits up to 150 sequences for SLAC analyses and 100 for FEL/IFEL analyses. As a rule of thumb, at least 10 sequences are needed to detect selection at a single site with any degree of reliability.

It is a good practice to visually inspect your data to make sure that the sequences are aligned correctly. Of course, one can never be sure that an alignment is objectively "correct", but gross misalignments (e.g. sequences that are out of frame) are easy to spot with software that provides a graphical visualization of the alignment, such as: MEGA, HyPhy, Se-Al or BioEdit. **You should verify that the alignment is in frame, i.e. that it does not contain stop codons, including premature stop codons, indicative of a frame shift, e.g. due to misalignment, or a non-functional coding sequence, and the terminal stop codon.** Your alignment should exclude any non-coding region of the nucleotide sequence, such as introns or promoter regions, for which existing models of codon substitution would not apply. When coding nucleotide sequences are aligned directly, frameshifting (i.e. not in multiples of 3) gaps may be inserted, since the alignment program often does not take the coding nature of the sequence into account. Therefore it is generally a good idea to align translated protein sequences and then map them back onto constituent nucleotides. Datamonkey will perform a number of checks when it receives coding sequences and report all problems it encounters.

If the alignment contains identical sequences, Datamonkey will discard all but one of the duplicate sequences before proceeding. This is done to speed up the analyses, because identical sequences do not contribute any information to the likelihood inference procedure (except via base frequencies), but the computational complexity of phylogenetic analyses grows with the number of sequences.

Finally, Datamonkey may rename some of the sequences to conform to HyPhy naming conventions for technical reasons (all sequence names must be valid identifiers, e.g. they cannot contain spaces). This is done automatically and has no effect on the subsequent analyses.

Common issues when preparing the data for Datamonkey.

Non-text files. Datamonkey expects sequence alignments to be uploaded as text files. Any other format (Word, RTF, PDF) will not be recognized and must be converted into plain text prior to submission.

Nonstandard characters in the alignment. For instance, BioEdit may use the tilde ('~') character to denote a gap. The dot ('.') character is sometimes used as 'match the first sequence' character and sometimes as the gap character. Datamonkey will accept IUPAC nucleotide characters (ACGT/U and ambiguity characters) and '?', 'X', 'N' or '-' for gap or missing data (Datamonkey is not case sensitive). All other characters in sequence data will be skipped and could result in frame shifts, which will be reported upon upload.

Uploading an amino-acid alignment. Datamonkey employs codon models which require the knowledge of silent substitutions, lost upon translation to amino-acids.

Termination codons. Datamonkey will reject any alignments that contains stop codons, even if the stop codon is at the end of the sequence (i.e. is a proper termination codon). Please strip all stop codons out of the alignment prior to up- loading it.

Alignments that are too gappy. If an alignment contains more than 50% of indels, it may not be properly processed (e.g. it could be read as a protein alignment, depending on the alignment format).

Alignments that are too large. If your alignment exceeds the size currently allowed by Datamonkey, consider running your analysis locally in HyPhy.

Incorrect genetic code. If the genetic code is misspecified (e.g. the mitochondrial code is applied to nuclear sequences), valid alignments may fail to up- load and if they do, then the results may be compromised (because codons are mistranslated). Make sure the correct genetic

Appendix 2: Copyright Attribution