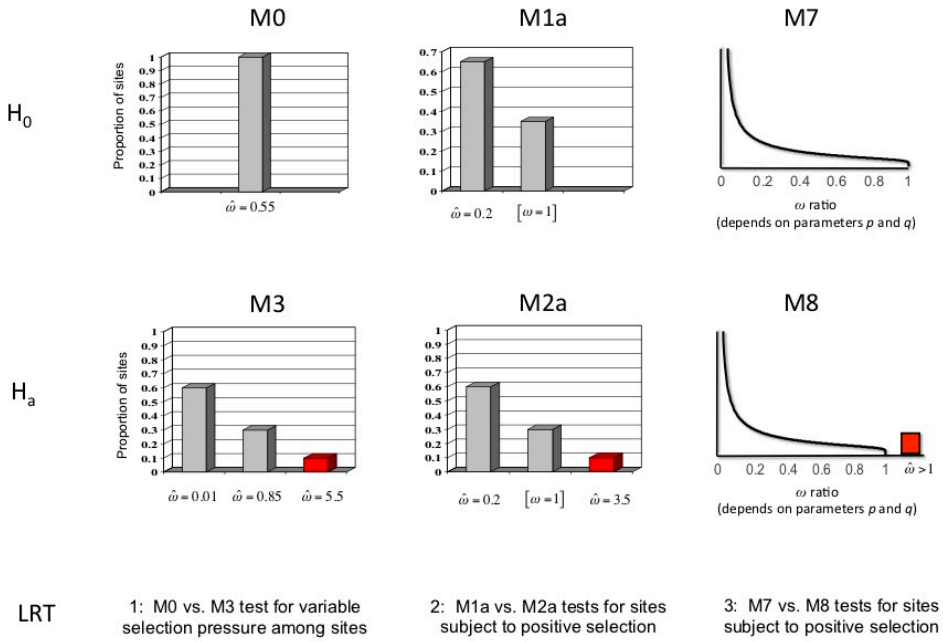
Exercise 4

The objective of this exercise is to use a series of LRTs to *test for sites* evolving under positive selection in the nef gene. If you find significant evidence for positive selection, then identify the involved sites by using empirical Bayes methods.

1. Obtain all the files for Exercise 4 (**ex4_codeml.ctl.txt**, **ex4_seqfile.txt**, **treeM0.txt, treeM1.txt, treeM2.txt, treeM3.txt, treeM7.txt, treeM8.txt**). When you are ready to run CODEML, remember to delete the **ex4**_ prefix (the control file must be called **codeml.ctl**).

2. If you plan to run two or more models at the same time, then create a separate directory for each run and place a copy of the sequence file, and the required control file and tree file in each directory.

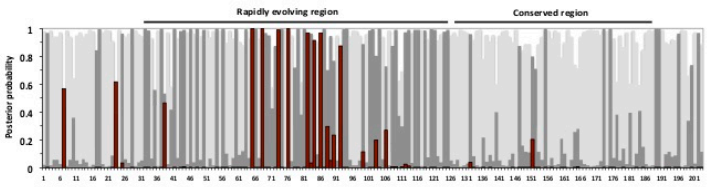


3. As in all the previous exercises, you will need to *change the control file* and re-run CODEML several times. In this case you will be fitting six different codon models (M0, M1a, M2a, M3, M7 & M8) to the example dataset. The control file "quick guide" might be helpful here (quick guide).

- If you are running your analyses sequentially in the same directory, then you should change the name of the main result file (via outfile= in the control file) or you will overwrite your previous results.
- Set the tree file with `treefile=`. I have supplied tree files pre-loaded with the ML branch lengths for each model (hence you need to set a different tree for each model). This will greatly speed up your analyses, giving you more "beer time". See the example control file for more details about treefile names\
- Set the codon model with `NSsites=`.
- Fix the value of kappa at the ML estimate with `kappa=`. Again, this will help speed up the analysis. See the control file for the value of kappa for each model.
- For some models you will also need to set the number of categories (ncatG) in the ω distribution:
  - for M3 set `ncatG=3`
  - for M7 set `ncatG=10`
  - for M8 set `ncatG=10`
- Once the analysis is complete, rename the rst file because subsequent runs will overwrite it!
- Repeat steps for each of the six codon models listed above.

4. Keep track of your results (ex4_HelpFile.pdf) by using a table like **Table E4** shown in the slides (see ex4 table template.pdf).
5. In addition, carry out the following likelihood ratio tests:
   - M0 vs. M3 (4 degrees of freedom)
   - M1a vs. M2a (2 degrees of freedom)
   - M7 vs. M8 (2 degrees of freedom)
6. Lastly, open the rst file generated when you ran model M3 (ex4_rst-HelpFile.pdf). Locate the columns of posterior probabilities for each site under the three site-categories of this model. Use these data to produce the plot for the nef gene like the one shown below (*your plot will look different than the one shown below*).

Example of a posterior distribution of selection pressure among sites



NOTE: This is **NOT** the distribution for the *nef* gene

7. As a final step, try to synthesize all your results and attempt a biological interpretation of the sort that you would want to publish within an actual research paper. The following two general questions should help get you going. I strongly encourage you to do this last step in collaboration with other workshop students; talk it through!

    - *What biological conclusions are well-supported by these data?*
    - *What aspects of the results can you interpret according your prior biological knowledge of this, or similar, systems?*