Exercise 4:

Testing for adaptive evolution in the *nef* gene of human HIV-2

Objectives: 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.

Step-by-step guide:

1. Obtain all the files for Exercise 4 from the course website.
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 sequence file, control file and tree file in each one.
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.
    a. 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.
    b. 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.
    c. Set the codon model with "`NSsites =`".
    d. 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.
    e. For some models you will also need to set the number of categories (ncatG) in the omega distribution:
        i. For M3 set "`ncatG = 3`"
        ii. For M7 set "`ncatG = 10`"
        iii. For M8 set "`ncatG = 10`"
    f. Once the analysis is complete, rename the "rst" file because subsequent runs will overwrite it!
    g. Repeat steps a. through f. for each of the six codon models listed above.
4. Keep track of your results by using a table like "Table E4" shown in the slides.
5. In addition, carry out the following likelihood ratio tests:
    a. M0 vs. M3 (4 degrees of freedom)
    b. M1a vs. M2a (2 degrees of freedom)
    c. M7 vs. M8 (2 degrees of freedom)
6. Lastly, open the "rst" file generated when you ran model M3. Locate the columns of posterior probabilities for each site under the three site-categories of this model. Use these data to reproduce the plot shown in the slides.