WORKED EXAMPLE 1

To demonstrate the use of the clade models in the preset running mode of EasyCodeML, we present an analysis of the ECP-EDN gene family based on data from a study by Bielawski and Yang (2003). The aim of this study is to investigate the role of positive selection in the ECP-EDN gene family in primates.

1. Choosing a running mode and input data

EasyCodeML has two different running modes, preset and custom. In this case, we choose the preset mode (Fig. 2A). We either drag-and-drop a folder into EasyCodeML or click on the button '…' to browse and select a local folder as working directory. The required inputs for analysing selection are the aligned sequences in PAML format and a tree file in Newick format. We can also drag-and-drop these two files into the text box. Four different model approaches are available in the preset mode. Here we select 'Clade Model' to test for positive selection in the ECP-EDN gene family (Fig. 2A).

2. Configure the running parameters

Once the sequence and tree files are selected, press the "Check" button to check the consistency of the taxon labels between the tree and sequence files. The clade models require the nodes of the tree to be labeled in order to indicate the clades that will be assigned independent omega parameters, so we press the "Label" button (Fig. 3A). We then click on the entire EDN clade to be selected in the tree as the foreground lineage. The dollar symbol '$' with an integer will be shown above the EDN clade. In EasyCodeML, the symbols '#' and '$' are used for the branch or branch-site models and for the clade model, respectively.

We use other default parameters, including the "Num of Threads" and "Clean data" options. Multithreading will only take effect on the site model. If the "Clean data" option is enabled, all sites with ambiguity characters and alignment gaps will be removed from the aligned sequence file prior to analysis.

3. CodeML analysis

Before starting the CodeML analysis, we need to click on the "Save Current Profile" button to enable all parameters for the current analysis. We then click on the button "Run CodeML" to start a CodeML analysis. Once the analysis is done, the log-likelihood (lnL) values and the number of parameters (np) will be automatically retrieved. A likelihood-ratio test is performed for the nested models and all results are automatically organized and displayed on the screen (Fig. 2A).

4. Summarizing and interpreting results

A publication-quality table that contains all of the relevant information from the CodeML analyses can be generated using the "Export" button. The "Microsoft Excel will be launched to view the saved results file by clicking the "View" button. A clear rejection of the null model indicates that divergent selection was detected between the foreground (the entire EDN clade) or background branches (the entire ECP clade). Note that the selection analysis presented here is merely instructional. To find the globally optimal likelihood score, we can load and edit the control file under the custom mode in EasyCodeML, and then run the program several times using different initial values of omega.