# **EasyCodeML:** an interactive visual tool for CodeML analysis

Fangluan Gao[1] & Chengjie Chen[2]

[1] Institute of Plant Virology, Fujian Agriculture and Forestry University

[2] College of Horticulture, South China Agricultural University

EasyCodeML was written in JAVA, available freely for non-commercial purposes at https://github.com/BioEasy/EasyCodeML.

# Contents


**1. INTRODUCTION** ..... 2

*1.1 Function* ..... 2

*1.2 Supported Platforms* ..... 2

**2. LICENSE & DISCLAIMER** ..... 2

*2.1 License* ..... 2

*2.2 Disclaimer* ..... 2

**3. SYSTEM REQUIREMENTS** ..... 2

**4. DATA PREPARATION** ..... 3

*4.1 Sequence File* ..... 3

*4.2 Tree File* ..... 3

*4.3 Labeling branches (or nodes) in a tree* ..... 3

**5. HOW TO RUN EASYCODEML** ..... 3

*5.1 The preset mode* ..... 3

*5.2 The custom mode* ..... 6

*5.3 EasyCodeML Menu* ..... 7

**6. EasyCodeML FAQ** ..... 8

*6.1 How to modify the control file under the preset mode?* ..... 8

*6.2 How to perform the likelihood ratio tests in EasyCodeML?* ..... 8

*6.3 How to increase memory to run EasyCodeML?* ..... 9

**7. ACKNOWLEDGMENTS** ..... 9

**8. REPORTING BUGS AND FEEDBACK** ..... 9

**9. APPENDIX** ..... 9

*9.1 How to convert other formats to PAML (*.PML)?* ..... 9

*9.2 How to convert a Tree to Newick format?* ..... 10

*9.3 Why can I not open EasyCodeML?* ..... 10

**10. REFERENCES** ..... 10

## 1. INTRODUCTION

EasyCodeML is an application to simplify the use of CodeML (Yang, 2007) by providing a user-friendly GUI.

### *1.1 Function*

- EasyCodeML provides two running modes, one for beginners and the other for experienced users;
- Nested codon models with predefined parameters were built in the preset mode, evolutionary analysis can be carried out with just one click；
- Single codon mode with self-definable parameters was provided in the custom mode to meet requirements of different data type;
- EasyCodeML provides a GUI with which tree labeling can be conducted more conveniently with increased accuracy (For both modes).
- EasyCodeML provides an “Export” module with which a publication-quality table could be generated by one click (Only for the preset mode);
- EasyCodeML supports multi-thread computation that significantly accelerate the phase of analysis and could be of much more convenience and efficiency. Precompiled EasyCodeML for multiple opening systems are also implemented

### *1.2 Supported Platforms*

EasyCodeML was written in Java. Precompiled EasyCodeML for multiple operation systems, including Microsoft Windows, Mac OS X and Linux, has been implemented.

## 2. LICENSE & DISCLAIMER

### *2.1 License*

EasyCodeML is a free software, and you are welcome to redistribute it under certain conditions. And it is released under GNU Lesser General Public License, Version 3. See http://www.gnu.org/licenses/lgpl.html

### *2.2 Disclaimer*

This program comes with ABSOLUTELY NO WARRANTY. No guarantee of the functionality of this software, or of the accuracy of results obtained, expressed or implied. Please inspect results carefully.

## 3. SYSTEM REQUIREMENT

EasyCodeML requires a **Java runtime environment** (JRE). Before running this program, **please make sure the correct installation of JRE 1.6** (or higher). Many systems will already have this installed. The latest versions of Java can be downloaded from http://www.java.com. If in doubt, type "java -version" in terminal (CMD in Windows, terminal in Mac or Linux)to check which version has been installed (Figure 1).

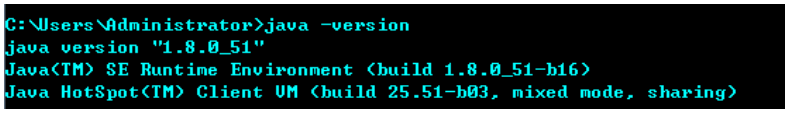


**Figure 1.** Using the Command-line to find Java Versions - Windows

## 4. DATA PREPARATION

### *4.1 Sequence File*

EasyCodeML fully supports PAML format (i.e., examples/example.nuc, Li et al, 2016), so PAML file does not require any modification. See also the appendix "***How to convert other formats to PAML (*.PML)***".

### *4.2 Tree File*

Tree file must be in Newick format (i.e., examples/example.trees, Li et al, 2016). See also the appendix "***How to convert a Tree to Newick format***" later about converting a tree format. **Note that taxon names with illegal characters (such as spaces, semicolons) are not allowed and the pairs of parentheses must be properly nested**.

### *4.3 Labeling branches (or nodes) in a tree*

Users are required to label branches (or nodes) in the tree for the branch (Yang 1998), the branch-site (Yang & Nielsen 2002; Yang et al. 2005; Zhang et al 2005) or the clade models (Weadick et al, 2012). For this purpose, EasyCodeML implemented a visualization interface enabling an interactive labeling of branches (or nodes) in a tree.

For the clade model, branch types (**NO MORE THAN FIVE**) are specified simply using the symbol "$" to label the clades in the tree file. For example, if you have five branch types, the symbols $1, $2, $3, $4 and $5 will successively label each clade. For the branch or branch-site models, branch labels are specified by the symbol "#" (fixed with the default label #1) to label the special branch in the tree file.

Certainly, user may manually label the branches or clades in the tree. And they can check whether the tree and the labeled branches (or clades) are correct using the "Label" module. See also the option "**Label**" later about interactive branch (or clade) labeling tree.

## 5. HOW TO RUN EASYCODEML

EasyCodeML provides two running modes, one for beginners (the preset mode, nested codon models with predefined parameters) and the other, for experienced users (the custom mode, independent codon models with user-defined parameters), respectively.

### *5.1 The preset mode*

In the preset mode, evolutionary analysis can be carried out with just one click because all key parameters of the nested models, including the branch model (M0 *vs.* two ratio Model 2), the branch-site model (ModelA null *vs.* Model A), the site model (M0 *vs.*M3, M1a *vs.* M2a and M7 *vs.* M8) and the clade model (M2a_rel *vs.* CmC), were built in and CodeML analysis from data input to results output was pipelined.

- **Working Directory:**
  Choose the directory where the data file should be saved.
- **Model Selection:**
  A model for CodeML analysis must be selected. Four types of codon models are available in EasyCodeML, including the branch model (M0 *vs.* two ratio Model 2), the branch-site model (ModelA null *vs.* Model A), the site model (M0 *vs.* M3, M1a *vs.* M2a and M7 *vs.* M8) and the clade model (M2a_rel *vs.* CmC).

- **Aligned Sequence File in PAML Format:**
  The aligned sequence file must be selected. Click on the button '…' and choose the file to be analyzed with PAML format or drag and drop directly the file to the textbox. Note that EasyCodeML does not check if the selected sequence file is of the right format.
- **Clean data:**
  If the option "Clean data" was selected, all sites with ambiguity characters and alignment gaps will be removed from the aligned sequence file. Note that alignment gaps are treated as missing data.

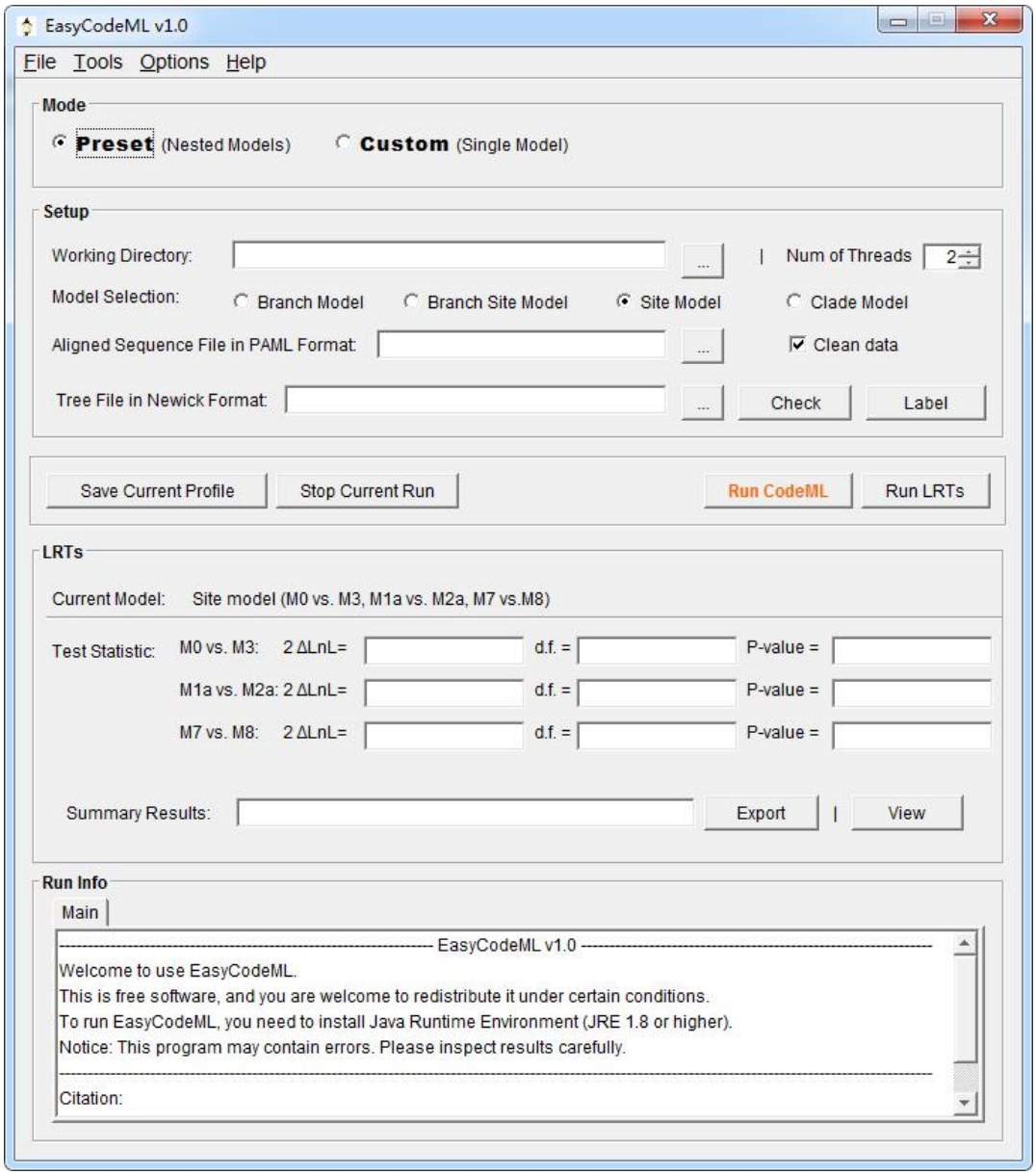


**Figure 2**. The main interface of EasyCodeML under the preset mode

- **Tree File in Newick Format:**
  The tree file must be selected as well. Click on the button '…' and choose the file to be analyzed with the Newick format. Similarly, note that EasyCodeML does not check if the selected tree file is of the right format.
- **Check:**

The taxa name discrepancies between the tree and sequence files will prevent user from running CodeML. To check these discrepancies, press the "Check" button when the sequence and tree files were selected. If there are no taxa name discrepancies in the data file (the sequence name in the aligned sequence file and the taxa name in the tree file is identical), a message box will be popped to notify that the user the data have been ready. If there are taxa name discrepancies in the data file, users should check them through the prompt box.

- **Label:**
  Press the "Label" button after a standard tree file has been selected, a dialog box for labeling tree appears. Then, click on the branch to be selected in the tree, the label marker (# or $) will be shown above the branch. When finish, click on the "Done" button. If you want to change the selected branch, just push the "Reset" button to reset it. If have multiple clades under the clade model, press the buttons 1st, 2nd, 3rd, 4th and 5th to label each clade (labeling $1, $2, $3, $4 and $5), respectively.

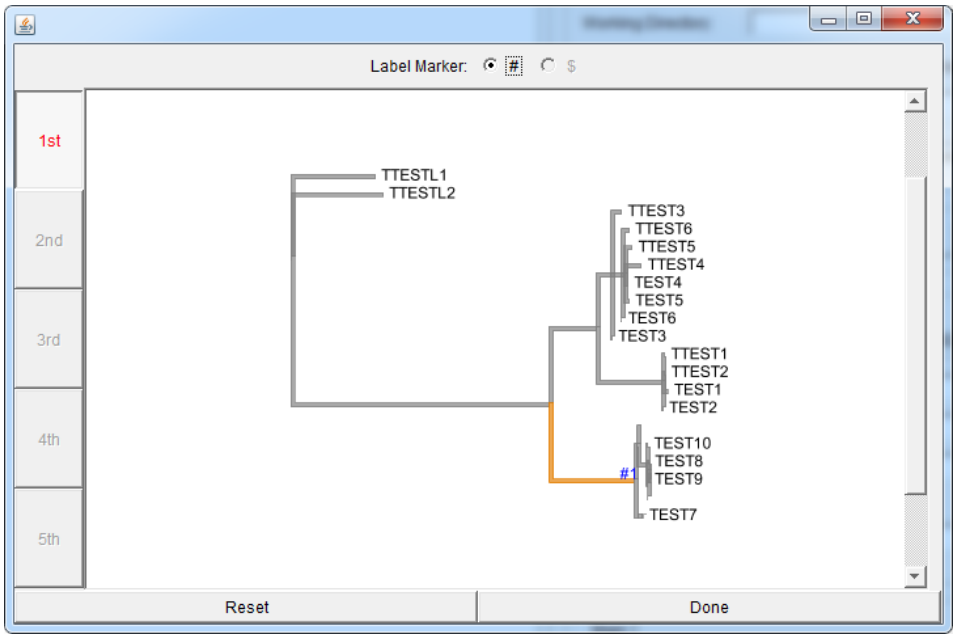


**Figure 3**. Labeling branches in a tree in a simpler and more intuitive way

- **Save Current Profile:**
  Click on the "Save Current Profile", the preset parameters for the current selected model are applied to CodeML analysis.
- **Run CodeML:**
  Click on the button "Run CodeML" to start a CodeML analysis and a running log in the main panel of "Run info" module shows live status of CodeML analysis.
- **Stop Current Run:**
  Click on the button "Stop Current Run" to stop the current CodeML analysis.
- **Run LRTs:**
  If a CodeML analysis is over, likelihood ratio tests (LRTs) will be done automatically and the results will be displayed on the screen. If users want to read the results from the last saved directory, Click the "Run LRT" button and be sure the Working Directory and Model Selection have been setup.

- **Export:**
  Click the “Export” button, the final results for the selected model can be summarized into a publication-quality table.
- **View:**
  Click the “View” button, Microsoft Excel will be launched and the saved results file will be opened.
- **Num of Threads:**
  The “Num of Threads” option will only take effect on the site model. The default of threads is 2, which can be specified according to the computer hardware configuration.

### *5.2 The custom mode*

Many experienced users need to create or modify the control files when using CodeML to meet different requirements. Therefore, the custom mode is implemented, in which all parameters for any single-model can be modified.

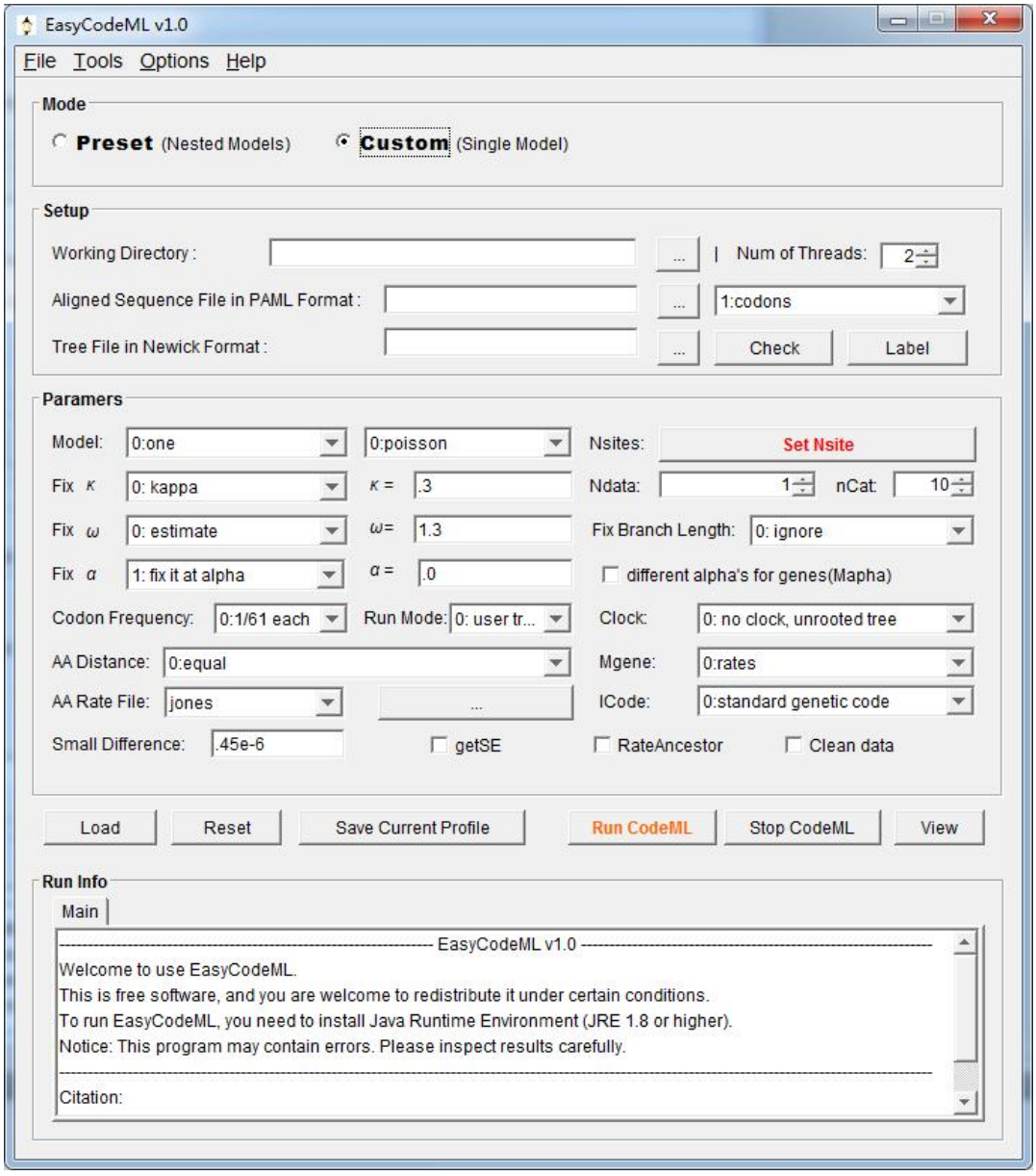

**Figure 4**. The main interface of EasyCodeML under the custom mode

- **Working Directory:**
  Choose the place where the data file should be saved.
- **Aligned Sequence File in PAML Format:**
  Click on the button '…' and choose the file to be analyzed with the PAML format or drag and drop directly the file to the textbox.
- **Tree File in Newick Format:**
  Click on the button '…' and choose the file to be analyzed with the Newick format.
- **Clean data:**
  If the option "Clean data" was selected, all sites with ambiguity characters and alignment gaps will be removed from the aligned sequence file. Note that alignment gaps are treated as missing data.
- **Check:**
  To check these discrepancies between the tree and sequence files, press the "Check" button when the sequence and tree files were selected.
- **Label:**
  Press the "Label" button to label branch (or clade) in the tree. Please see the previous section.
- **Load:**
  Click on the button "Load" to read and load the parameters from an existing CodeML control file (codeml.ctl).
- **Reset:**
  Click on the button "Reset" to reset the parameters showed on the main interface.
- **Save Current Profile:**
  Click on the "Save Current Profile", the current parameters are applied to CodeML analysis.
- **Run CodeML:**
  Click on the button "Run CodeML" to start a CodeML analysis and a running log in the main panel of "Run info" module shows live status of CodeML analysis.
- **Stop Current Run:**
  Click on the button "Stop Current Run" to stop the current CodeML analysis.
- **View:**
  Click on the button "View" to open the working directory of the current CodeML analysis.

For more information on the usage of the options within the parameter textbox, please refer to the PAML user manual.

PAML user manual is available from http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf

### *5.3 EasyCodeML Menu*

| Menu | Submenu | Description |
|---|---|---|
| File | | |
| | Load Aligned Sequence | Load a codon-based alignment |
| | Load Tree File | Load a tree file |
| | Exit | Quit the program |
| Tools | | |

| | | |
|---|---|---|
| | LRTs Calculator | Retrieve *p*-values for LRTs |
| | Control File Editor | Edit a CodeML control file |
| Options | | |
| | Configure Tree Label | Modify tree layout to fit in the display window |
| Help | | |
| | User guide | EasyCodeML user manual |
| | About | About EasyCodeML |

## 6. EasyCodeML FAQ

### *6.1 How to modify the control file under the preset mode?*

Under the preset mode, the control file for different codon models is implemented within the preset directory, user can load, modify and save the control file through the *Control file Editor* in the *Tools* menu. **Note that this should be done with caution because the parameters in the control file will be permanently modified.**

### *6.2 How to perform the likelihood ratio tests in EasyCodeML?*

Under the preset mode, likelihood ratio tests (LRTs) will be done automatically after a pair of nested models been tested.

LRT results also can be read from the last saved directory using "Run LRT" as shown as the following steps (Figure 5).

(1) Specify the last saved destination folder path to current working directory (i.e. E:\Biosofts\EasyCodeML\test).

(2) Choose the model for the last CodeML analysis.

(3) Click the "Run LRT" button, press the Y key after a dialog appears and the LRTs results will be displayed on the screen.

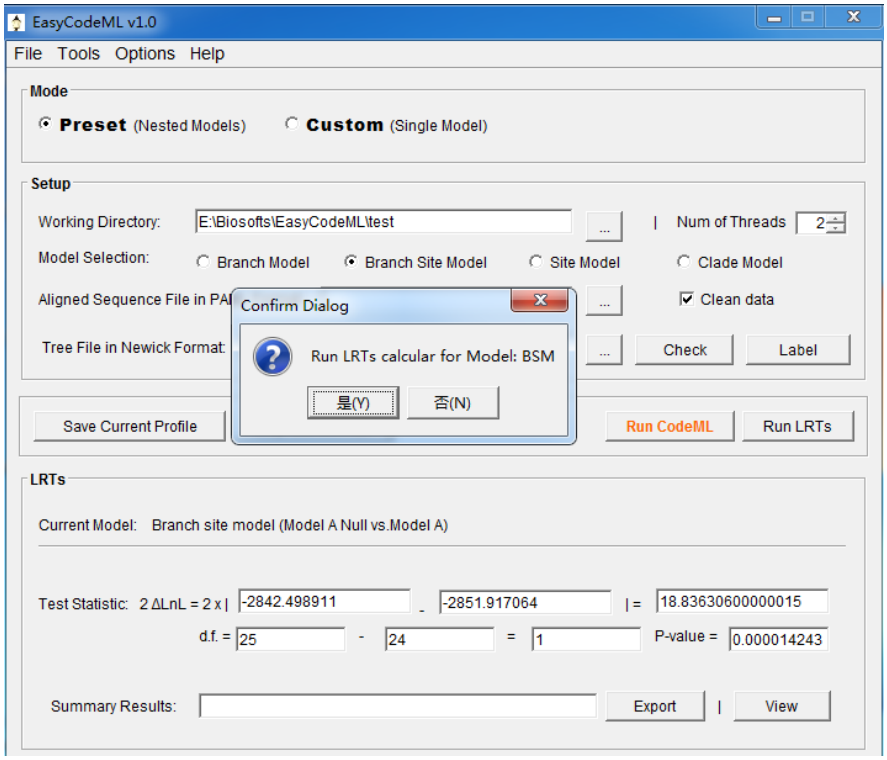

**Figure 5.** "Run LRT" module for reading LRT results from the last saved directory

Under the custom mode, LRTs can be conducted using the LRT calculator available in the program's menu as shown as the following steps (Figure 6).

(1) Input the log likelihood values of the nested model (LnL0 and LnL1) to get the value of 2ΔLnL.

(2) Input the np values of the nested models (np0 and np1) to get the degree of freedom (d.f.).

(3) Press the "Run LRTs" button, then we will get a p-value for this LRT.

(4) Press the "Copy" button, the p-value for this LRT be copied onto the clipboard.

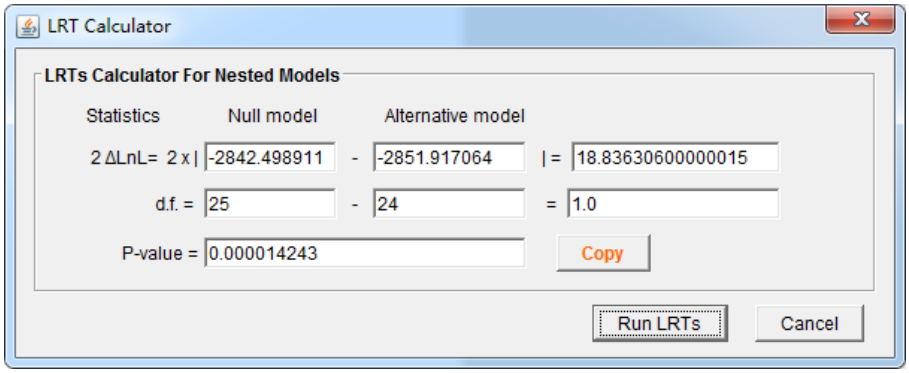


**Figure 6.** LRT Calculator in the **Tools** menu

### *6.3 How to increase memory to run EasyCodeML?*

To increase the memory to run EasyCodeML, it is allowed to start the program by executing the command "java -Xmx1024m -Xms512M -jar EasyCodeML.jar" and adapt the -Xmx parameter to your needs (-Xmx1024m means: maximum memory of 1,024 MB).

## 7. ACKNOWLEDGMENTS

We thank Mrs. Zhenxi Chen (Tropical crops genetic resources institute, CATAS), Han Li (Southwest University), Lin Zhang (Nanjing Normal University) for constructive feedback in the development of the program.

## 8. REPORTING BUGS AND FEEDBACK

If you encounter any problems with EasyCodeML, please let us know. We will be sure to promptly track down and fix any bugs that are reported. If it is possible, also attach the data which caused the problems.

EasyCodeML is an on-going project. For any suggestions of further features that you would like to see included in the next version, please let me know as well. We make no promises, however...

Also, if you find EasyCodeML useful then we would welcome any positive feedback or suggestions!!!

- Fangluan Gao [ raindy@fafu.edu.cn ]
- Chengjie Chen [ chengjiechen@stu.scau.edu.cn ]

## 9. APPENDIX

### *9.1 How to convert other formats to PAML (*.PML)?*

(1) Start DAMBE, and open a standard sequence file and choose the correct sequence type (i.e. Protein-coding Nuc. Seq). The sequences will be showed on the display windows. DAMBE is available from http://dambe.bio.uottawa.ca/DAMBE/dambe.aspx .

(2) Click **File | Save or Convert Sequence Format**, the standard File/Open dialog box appears. Choose the PAML file format and click OK. You will be informed that the file has been saved in a PAML file.

(3) Click **OK**, the converted file will be displayed on the screen.

### *9.2 How to convert a Tree to Newick format?*

(1) Start FigTree, and open a standard tree file. The tree file will be display in the main interface of FigTree. FigTree is available from http://tree.bio.ed.ac.uk/software/figtree/

(2) Click **File | Export Trees**, and select **Newick** from menu.

(3) Click **OK**, the tree will be stored in the Newick format.

### *9.3 Why cannot I open EasyCodeML?*

(1) Make sure the installation of JRE 1.6 (or higher) before running this program.

(2) You'll need to modify your default program configuration so that file extension .jar is not associated with other programs like WinRAR (Figure 7) and something else. Or you can re-associate .jar file with the JVM or use "open with option" in windows, or using command "java -jar EasyCodeML.jar" in your terminal (CMD in windows or Terminal in Mac or Linux) to start the program.

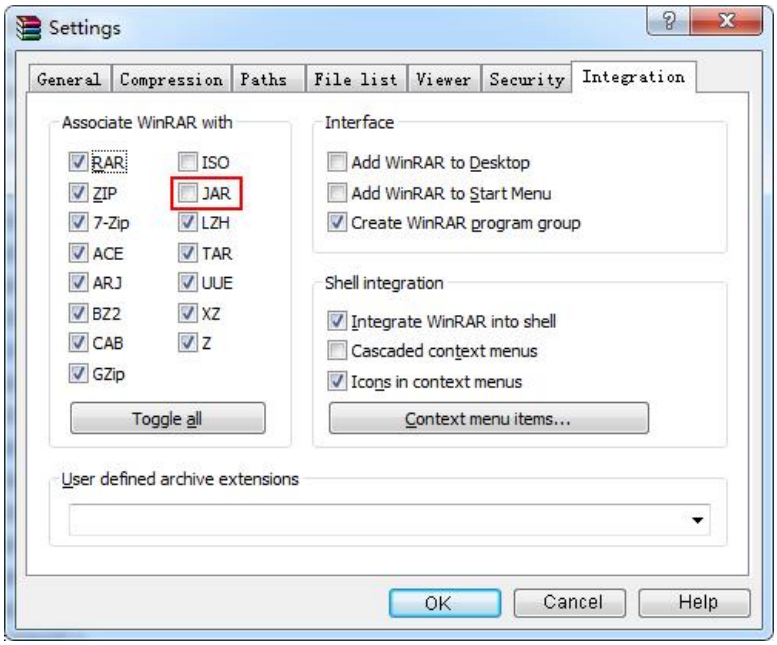


**Figure 7.** Make sure the file extension .jar not to be associated with WinRAR

## 10. REFERENCES


Li, H., *et al*. Evolutionary and functional analysis of mulberry type III polyketide synthases. *BMC Genomics* 2016;17(1):1-18.

Nielsen, R. and Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 1998;148(3):929-936.

Weadick, C.J. and Chang, B.S. Complex patterns of divergence among green-sensitive (RH2a) African cichlid opsins revealed by Clade model analyses. *BMC evolutionary biology* 2012;12(1):1-17.
Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution* 1998;15(5):568-573.
Yang, Z. and Nielsen, R. Codon-substitution models for detecting molecular adaptation at idividual sites along specific lineages. *Molecular biology and evolution* 2002;19(6):908-917.
Yang, Z., *et al*. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 2000;155(1):431-449.
Yang, Z., Wong, W.S.W. and Nielsen, R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular biology and evolution* 2005;22(4):1107-1118.
Zhang, J., Nielsen, R. and Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution* 2005;22(12):2472-2479.