PAMLX: A Graphical User Interface for PAML

Bo Xu[1,2] and Ziheng Yang*[1,2,3]

[1]Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[2]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[3]Department of Genetics, Evolution, and Environment, University College London, London, United Kingdom

***Corresponding author:** E-mail: z.yang@ucl.ac.uk.

Associate editor: Koichiro Tamura

Abstract

This note announces PAMLX, a graphical user interface/front end for the PAML (for Phylogenetic Analysis by Maximum Likelihood) program package (Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556; Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591). PAMLX is written in C++ using the Qt library and communicates with PAML programs through files. It can be used to create, edit, and print control files for PAML programs and to launch PAML runs. The interface is available for free download at http://abacus.gene.ucl.ac.uk/software/paml.html.

Key words: **phylogenetics, molecular evolution, software.**

PAML is a program package for phylogenetic analysis of DNA and protein sequences using maximum likelihood (ML) (Yang 1997, 2007). It includes the following programs: BASEML and BASEMLG for nucleotide-based substitution models, CODEML for amino acid and codon sequences, MCMCTREE for Bayesian estimation of species divergence times, YN00 for estimation of d_N and d_S by the method of Yang and Nielsen (2000), as well as the simulation program EVOLVER. The strength of the package lies in its implementation of a rich collection of substitution models, useful for estimating parameters in models of sequence evolution and for testing various hypotheses. Examples of analyses that can be performed using the package include likelihood-based tests of phylogenetic trees (BASEML and CODEML) (Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999); estimation of parameters in models of variable rates among sites and models for combined analysis of multiple genes (BASEML and CODEML) (Yang 1994, 1996); likelihood ratio tests of various hypotheses through comparison of nested statistical models (BASEML and CODEML) (Yang 2006, Chapters 1, 2, and 4); estimation of d_N and d_S and detection of positive selection in protein-coding genes (YN00 and CODEML) (Yang 2006, Chapter 8); estimation of

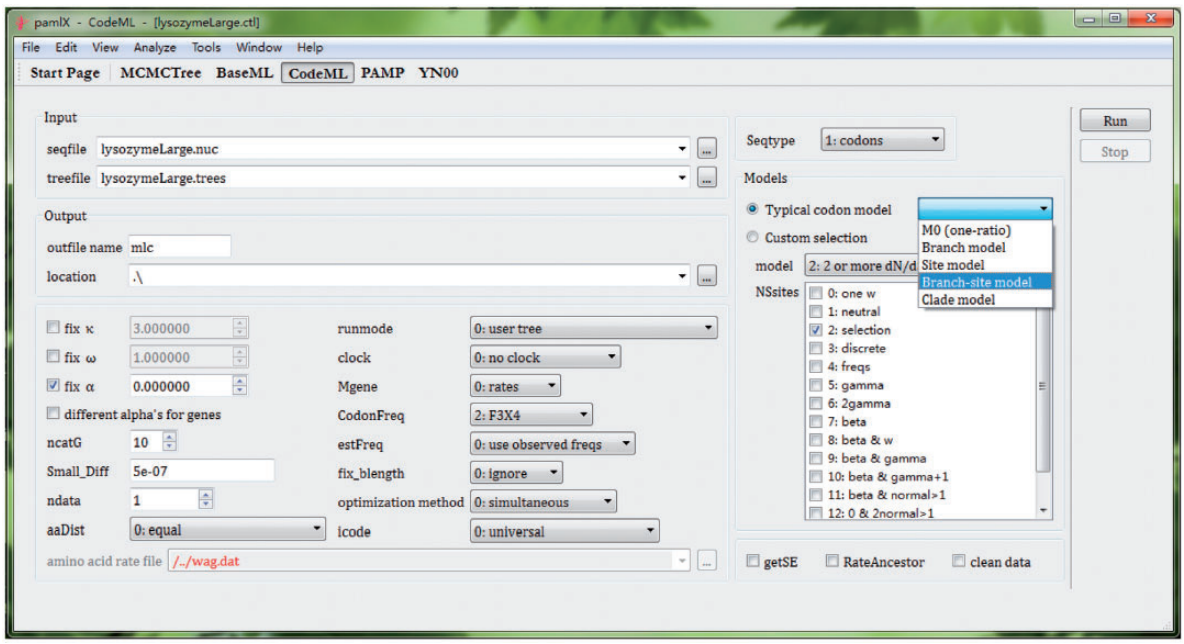

FIG. 1. Screen shot for PAMLX interface for the CODEML program, highlighting a list of commonly used codon models.

empirical amino acid substitution matrices (CODEML) (Yang et al. 1998); Bayesian estimation of species divergence times under clock and relaxed-clock models (MCMCTREE) (Yang and Rannala 2006; Rannala and Yang 2007); reconstruction of ancestral sequences using nucleotide, amino acid, and codon models (BASEML and CODEML) (Yang et al. 1995); and generation of nucleotide, codon, and amino acid sequence alignments by Monte Carlo simulation (EVOLVER). For an overview of the package, see Yang (2007).

The PAML programs use a simple command-line interface, and the task of preparing the control files in a text editor may appear daunting to beginning users. We have thus developed a graphical user interface (GUI) for the package, called PAMLX, which can also be used as a front-end to launch PAML analyses. The main function of the interface is to create, read, edit, save, or print the control files for the main programs in the PAML package, including BASEML, CODEML, and MCMCTREE. For example, a number of codon-based models have been implemented in the CODEML program, including the branch model for detecting lineages under positive selection (Yang 1998), the site models for detecting individual amino acid sites under positive selection (Nielsen and Yang 1998; Yang et al. 2000, 2005), the branch-site models for detecting positive selection affecting a few sites on particular lineages, and the clade models for detecting functional divergence among clades (clusters of branches) (Yang and Nielsen 2002; Zhang et al. 2005; Weadick and Chang 2012). Specification of those models is now simplified by having a pull-down menu listing the commonly used codon models (fig. 1).

The PAMLX interface is written using the Qt cross-platform application framework. Compiled executables for Windows, Linux, and Mac OSX are provided. Communication with the PAML programs is through files, and the location of the PAML package is specified inside the interface. A benefit of this design is that future updates of the PAML package will require no or minimal changes to the interface. However, this design has the drawbacks that the two pieces of software are only loosely connected. For example, PAMLX is no more flexible about sequence data formats than is PAML and only phylip and Nexus formats are accepted, and there is no effective method for handling errors from PAML programs. Currently, PAMLX can run only one PAML job at a time. Future improvements may include managing multiple PAML jobs on multicore multiprocessor machines.

Acknowledgments

The authors thank Dr Fengrong Ren and two anonymous reviewers for many useful suggestions, Jose Barba for testing the Windows version, and Ying Wang for testing the Mac OSX version of the software. This work was supported by a BBSRC grant to Z.Y.

References

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol.* 29:170–179.

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

Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol.* 56:453–466.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol.* 16:1114–1116.

Weadick CJ, Chang BS. 2012. An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. *Mol Biol Evol.* 29:1297–1300.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39:306–314.

Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol.* 42:587–596.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 15: 568–573.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol.* 19:908–917.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol.* 23:212–226.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.

Yang Z, Nielsen R, Goldman N, Pedersen A-MK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol.* 15:1600–1611.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22: 1107–1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.