PAML 4: Phylogenetic Analysis by Maximum Likelihood

Ziheng Yang*

*Department of Biology, Galton Laboratory, University College London, London, United Kingdom

PAML, currently in version 4, is a package of programs for phylogenetic analyses of DNA and protein sequences using maximum likelihood (ML). The programs may be used to compare and test phylogenetic trees, but their main strengths lie in the rich repertoire of evolutionary models implemented, which can be used to estimate parameters in models of sequence evolution and to test interesting biological hypotheses. Uses of the programs include estimation of synonymous and nonsynonymous rates ($d_S$ and $d_K$) between two protein-coding DNA sequences, inference of positive Darwinian selection through phylogenetic comparison of protein-coding genes, reconstruction of ancestral genes and proteins for molecular restoration studies of extinct life forms, combined analysis of heterogeneous data sets from multiple gene loci, and estimation of species divergence times incorporating uncertainties in fossil calibrations. This note discusses some of the major applications of the package, which includes example data sets to demonstrate their use. The package is written in ANSI C, and runs under Windows, Mac OSX, and UNIX systems. It is available at http://abacus.gene.ucl.ac.uk/software/paml.html.

Introduction

Phylogenetic methods for comparative analysis of DNA and protein sequences are becoming ever more important with the rapid accumulation of molecular sequence data, spearheaded by numerous genome projects. It is now common for phylogeny reconstruction to be conducted using large data sets involving hundreds or even thousands of genes. Similarly, phylogenetic methods are widely used to estimate the evolutionary rates of genes and genomes to detect footprints of natural selection, and the evolutionary information is used to interpret genomic data (Yang 2005). For example, both evolutionary conservation indicating negative purifying selection and accelerated evolution driven by positive Darwinian selection have been employed to detect functionally significant regions of the genome (e.g., Thomas et al. 2003; Nielsen et al. 2005; Sawyer et al. 2005).

PAML is a package of programs for phylogenetic analyses of DNA and protein sequences using maximum likelihood (ML). The package includes the following programs: BASEML, BASEMLG, CODEML, EVOLVER, PAMP, YN00, MCMTREE, and CHI2. Tree-search algorithms implemented in BASEML and CODEML are primitive. However, the programs may be used to evaluate a collection of trees obtained using other programs such as PHYLIP (Felsenstein 2005), PAUP (Swofford 2000), MRBAYES (Huelsenbeck and Ronquist 2001) and MEGA (Kumar, Tamura, and Nei 2005). The strength of PAML is in its rich collection of sophisticated substitution models, useful when our focus is on understanding the process of sequence evolution. Examples of analyses that can be performed using the package include:

- Comparison and tests of phylogenetic trees (BASEML and CODEML);
- Likelihood ratio tests (LRTs) of hypotheses through comparison of nested statistical models (BASEML, CODEML, CHI2);
- Estimation of synonymous and nonsynonymous substitution rates and detection of positive Darwinian selection in protein-coding DNA sequences (YN00 and CODEML);
- Estimation of empirical amino acid substitution matrices (CODEML);
- Estimation of species divergence times under global and local clock models using likelihood (BASEML and CODEML) and Bayesian (MCMTREE) methods;
- Reconstruction of ancestral sequences using nucleotide, amino acid, and codon models (BASEML and CODEML);
- Generation of nucleotide, codon, and amino acid sequence alignments by Monte Carlo simulation (EVOLVER).

This article provides an overview of a few major applications of PAML programs, with an emphasis on models and analyses in common use but unavailable elsewhere. Example data files used in publications that described those methods are included in the package, to illustrate the file formats and the interpretation of results (see table 1). New users of the programs are advised to use the examples to duplicate published results before analyzing their own data.

Major Applications of the Software Package

Comparison and Tests of Trees

The programs BASEML and CODEML can take a set of user trees and evaluate their log likelihood values under a variety of nucleotide, amino acid, and codon substitution models. When more than one tree is specified, the programs automatically calculates the bootstrap proportions for trees using the RELL method (Kishino and Hasegawa 1989), as well as $p$ values using the K-H test (Kishino and Hasegawa 1989) and S-H test (Shimodaira and Hasegawa 1996). See Goldman, Anderson, and Rodrigo (2000) for a critical review of those methods.

In particular, a number of likelihood models are implemented in BASEML and CODEML for combined analysis of heterogeneous data sets from multiple gene loci (Yang 1996). These models allow estimation of common parameters
of interest (such as species divergence times or species phylogenies), while allowing other parameters (such as the substitution rate, transition/transversion rate ratio, and base compositions) to differ among loci to accommodate the idiosyncrasies in the evolutionary process at different loci. Application of such models in the ML method has been discussed by Yang (1999), Pupko et al. (2002), Shapiro, Rambaut, and Drummond (2006), and Bofkin and Goldman (2007). Similar models are implemented in the Bayesian framework by Suchard et al. (2003) and Nylander et al. (2004). The use of statistical likelihood in those methods allows one locus to borrow information from other loci in combined analysis of heterogeneous data sets, and the method combines the strengths of the supermatrix and supertree methods while avoiding their drawbacks.

Estimation of Synonymous and Nonsynonymous Rates Between Two Protein-Coding DNA Sequences

In analysis of protein-coding genes, we have the advantage of being able to distinguish the synonymous or silent substitutions (nucleotide substitutions that do not change the encoded amino acid) from the nonsynonymous or replacement substitutions (those that do change the amino acid). Because natural selection operates mainly on the protein level, synonymous and nonsynonymous mutations are under very different selective pressures and are fixed at very different rates. Thus comparison of synonymous and nonsynonymous substitution rates can reveal the direction and strength of natural selection acting on the protein (Kimura 1977; Miyata and Yasunaga 1980).

The \( \text{YN00} \) program implements a number of counting methods for estimating \( d_0 \) and \( d_2 \) between two sequences, including NG86 (Nei and Gojobori 1986), LWL85 (Li, Wu, and Luo 1985), LPB (Li 1993; Pamilo and Bianchi 1993), and LWL85m, a modified version of LWL85 (Yang 2006, p. 55). NG86 assumes no transition-transversion rate difference and no codon usage bias, while LWL85, LPB, and LWL85m all account for the transition-transversion rate difference to some extent but assume no codon usage bias. The counting method of Yang and Nielsen (2000) is implemented in \( \text{YN00} \) as well, which accommodates both the transition-transversion rate difference and unequal codon frequencies. The ML method of Goldman and Yang (1994) is implemented in CODEML, which can be applied under different model assumptions. It can also be used to calculate ML estimates of a few new distances defined by Yang (2006; section \( \S 2.5.4 \)). These include \( d_{3B} \), the distance at the third codon position before the operation of natural selection on the protein level, which is very similar to \( d_4 \), the sequence distance at the so-called 4-fold degenerate sites. Usually a third codon position is considered to be a 4-fold site if the first and second positions are identical across sequences and if the encoded amino acid does not depend on the third position (e.g., Adachi and Hasegawa 1996). This definition has the drawback that the number of 4-fold sites used in the calculation decreases with sequence divergence. In CODEML, \( d_4 \) is calculated from the codon model and avoids this drawback (Yang 2006, p. 64).

Detection of Adaptive Molecular Evolution Under Models of Codon Substitution

A number of codon substitution models have been implemented in CODEML as extensions of the basic model of Goldman and Yang (1994; see also Muse and Gaut 1994). Several reviews have been published that discuss applications of those models to detect positive selection (Yang and Bielawski 2000; Nielsen 2001; Yang 2002, 2006, Chapter 8).

The main feature of codon-substitution models, compared with models of nucleotide or amino acid substitution, is that the codon triplet is considered the unit of evolution (Goldman and Yang 1994). The version of the model in common use (e.g., Yang 1998; Yang and Nielsen 1998) ignores chemical differences between amino acids and uses the same nonsynonymous/synonymous rate ratio \( \omega = d_0/d_2 \) for all nonsynonymous substitutions. This assumption may be unrealistic but simplifies the interpretation of the model. The \( \omega \) ratio measures the direction and magnitude of selection on amino acid changes, with values of \( \omega < 1, = 1, \) and \( > 1 \) indicating negative purifying selection, neutral evolution, and positive selection, respectively. However, straightforward use of the \( \omega \) ratio to detect positive selection, by calculating \( d_0 \) and \( d_2 \) between two sequences, is rarely effective, because the \( \omega \) ratio averaged over all sites is seldom greater than 1. Much effort has been taken recently to develop models useful for detecting positive selection that affects specific lineages or individual sites.

The branch models use different \( \omega \) ratio parameters for different branches on the phylogeny (Yang 1998; Yang 2002, 2006, Chapter 8);
and Nielsen 1998). They may be used to detect positive selection acting on particular lineages, without averaging the $\omega$ ratio throughout the phylogenetic tree. The branch models are useful for detecting positive selection after gene duplications, where one copy of the duplicates may have acquired a new function and may have thus evolved at accelerated rates.

The site models treat the $\omega$ ratio for any site (codon) in the gene as a random variable from a statistical distribution, thus allowing $\omega$ to vary among codons (Nielsen and Yang 1998; Yang et al. 2000). Positive selection is defined as presence of some codons at which $\omega > 1$. An LRT is constructed to compare a null model that does not allow for any codons with $\omega > 1$ against a more general model that does. Real data analyses and computer simulations (Anisimova, Bielawski, and Yang 2001, 2002; Anisimova, Nielsen, and Yang 2003; Wong et al. 2004) suggest that two pairs of site models are particularly effective (table 2). The first pair include M1a (neutral) and M2a (selection) (Nielsen and Yang 1998; Wong et al. 2004; Yang, Wong, and Nielsen 2005), while the second pair include M7 (beta) and M8 (beta&$\omega$) (Yang et al. 2000). The LRT statistic, or the log likelihood difference between the two compared models (2$\Delta$ln), may be compared against $\chi^2_{1}$ (with critical values to be 2.71 and 5.41 at the 5% and 1% significance levels, respectively) (Self and Liang 1987). Some authors (e.g., Zhang, Nielsen, and Yang 2007) also suggested the use of $\chi^2_{1}$ (with critical values to be 3.84 and 5.99), to guide against violations of model assumptions. The BEB method is implemented to calculate posterior probabilities for site classes under model A if the LRT suggests presence of codons under positive selection on the foreground branch.

Reconstruction of Ancestral Sequences

The programs BASEML and CODEML implement the empirical Bayes (EB) method for reconstructing genes or proteins in extinct ancestors on a phylogeny (Yang, Kumar, and Nei 1995; Koshi and Goldstein 1996). The general methodology is the same as the NEB for detecting positively selected sites discussed above. Ancestral reconstruction can be conducted under nucleotide-, amino acid-, or codon-based models. Compared with the parsimony algorithm (Fitch 1971; Hartigan 1973), the EB approach takes into account differences in the branch lengths and in the relative substitution rates between characters (nucleotides, amino acids, or codons). Both marginal and joint reconstructions are implemented. The marginal reconstruction assigns a character state to a single node on the tree and may be more suitable when the gene or protein sequence in an extinct ancestor is desired, as in restoration studies (Chang, Kazmi, and Sakmar 2002; Thornton 2004). The joint reconstruction assigns a set of character states to all ancestral nodes on the tree and is more appropriate when one counts changes at every site (Pupko et al. 2000). The EB approach replaces parameters in the model such as branch lengths and substitution parameters by their ML estimates. An alternative method, the hierarchical (full) Bayesian approach, accommodates sampling errors in the parameter estimates by averaging the parameters of the space of model A, the test statistic 2$\Delta$ln should be compared with the 50:50 mixture of point mass 0 and $\chi^2_{1}$ (with critical values to be 2.71 and 5.41 at the 5% and 1% significance levels, respectively) (Self and Liang 1987).

### Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>NSites</th>
<th>$p$</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0 (one ratio)</td>
<td>0</td>
<td>1</td>
<td>$\omega$</td>
</tr>
<tr>
<td>M1a (neutral)</td>
<td>1</td>
<td>2</td>
<td>$p_0 (p_1 = 1 - p_0)$, $\omega_0 = 1$, $\omega_1 = 1$</td>
</tr>
<tr>
<td>M2a (selection)</td>
<td>2</td>
<td>4</td>
<td>$p_0 p_1 (p_2 = 1 - p_0 - p_1)$, $\omega_0 &lt; 1$, $\omega_1 = 1$, $\omega_2 &gt; 1$</td>
</tr>
<tr>
<td>M3 (discrete)</td>
<td>3</td>
<td>5</td>
<td>$p_0 p_1 (p_2 = 1 - p_0 - p_1)$, $\omega_0 = \omega_1$, $\omega_2$</td>
</tr>
<tr>
<td>M7 (beta)</td>
<td>7</td>
<td>2</td>
<td>$\omega = p q$</td>
</tr>
<tr>
<td>M8 (beta&amp;$\omega$)</td>
<td>8</td>
<td>4</td>
<td>$p_0 (p_1 = 1 - p_0)$, $\omega = p q$, $\omega_2 &gt; 1$</td>
</tr>
</tbody>
</table>

**Note:**—The site models are implemented using the control variable NSites in CODEML, and $p$ is the number of free parameters in the $\omega$ distribution.

### Table 3

<table>
<thead>
<tr>
<th>Site class</th>
<th>Proportion of sites</th>
<th>Background $\omega$</th>
<th>Foreground $\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$p_0$</td>
<td>$0 &lt; \omega_0 &lt; 1$</td>
<td>$0 &lt; \omega_0 &lt; 1$</td>
</tr>
<tr>
<td>1</td>
<td>$p_1$</td>
<td>$\omega_0 = 1$</td>
<td>$\omega_1 = 1$</td>
</tr>
<tr>
<td>2a</td>
<td>$(1 - p_0 - p_1)$</td>
<td>$0 &lt; \omega_0 &lt; 1$</td>
<td>$\omega_2 &gt; 1$</td>
</tr>
<tr>
<td>2b</td>
<td>$(1 - p_0 - p_1)$</td>
<td>$\omega_2 = 1$</td>
<td>$\omega_1 = 1$</td>
</tr>
</tbody>
</table>

**Note:**—Model A is the alternative hypothesis for the branch-site test of positive selection. The null model fixes $\omega_1 = 1$ for the foreground branch.
over a prior (Huelsenbeck and Bollback 2001). This difference between methods may be important in small data sets that lack information to estimate the parameters reliably.

Ancestral reconstruction provides an intuitive way of exploring the data. It has been used in numerous analyses, for example, to estimate selective pressures along lineages (Messier and Stewart 1997; Zhang, Kumar, and Nei 1997) or at individual sites (Suzuki and Gojobori 1999). However, the intuitive simplicity of the idea invites its misuse. Most studies making use of ancestral reconstructions ignore the fact that the ancestral sequences are inferred pseudo-data instead of real observed data, and that using only the optimal character states while ignoring the suboptimal states can lead to systematic biases (Eyre-Walker 1998; Yang 2006, section §4.4).

Estimation of Species Divergence Times

The programs BASEML and CODEML implement the likelihood method of Yoder and Yang (2000; see also Yang and Yoder 2003) for estimating species divergence times under local-clock models. This, like the quartet-dating method of Rambaut and Bromham (1998), assigns rates to different branches on the tree and then estimates both branch rates and divergence times from the sequence data. The method allows analysis of data from multiple loci, taking into account differences in their evolutionary dynamics, and can use multiple fossil calibrations in the same analysis. However, the assignment of rates to branches in the method is somewhat arbitrary. A rate-smoothing procedure was implemented to help assign rates to branches automatically (Yang 2004), which may be considered an improved version of Sanderson’s (2002) penalized likelihood method.

A problem with current likelihood approaches to divergence time estimation is that fossil calibrations are assumed to provide known node ages without errors. The methods thus fail to account for uncertainties in fossil calibrations. Attempts to incorporate fossil uncertainties in the likelihood framework have not been very successful (see Yang 2006, section §7.3).

The program MCMCTREE used to implement the Bayesian method of phylogeny reconstruction of Yang and Rannala (1997; see also Rannala and Yang 1996). It was very slow and was decommissioned after the release of MRBAYES (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). The program now implements the MCMC algorithm for dating species divergences of Yang and Rannala (2006) and Rannala and Yang (2007). This is similar to the MULTIDIVTIME program (Thorne, Kishino, and Painter 1998; Kishino, Thorne, and Bruno 2001) and can be used to analyze data of multiple gene loci, incorporating multiple fossil calibrations and allowing the evolutionary rate to drift over time. A difference that may be important is that MULTIDIVTIME uses “hard” lower and upper bounds on node ages to specify fossil calibration information, while MCMCTREE uses “soft” bounds or arbitrary statistical distributions. A detailed discussion between the two programs is provided by Yang (2006, pp. 245–257).

Simulating Molecular Evolution

The program EVOLVER can be used to simulate DNA and protein sequences under various nucleotide, amino acid, and codon substitution models. Simulation is useful for learning about a complicated model or method of data analysis, and for comparing different analytical methods. For example, a number of simulation studies have been conducted to compare different tree reconstruction methods. In a simulation, the true model and the true values of parameters are under the control of the investigator, and they can be varied to examine their effects on method performance. See Yang (2006, Chapter 10) for a description of techniques for simulating molecular evolution.

Software Platform and Availability

The programs in the PAML package are written in ANSI C. Executables are compiled for Windows and Mac OSX, and C source codes are provided for UNIX systems. The program is distributed free of charge for academic use at its web site: http://abacus.gene.ucl.ac.uk/software/paml.html.

Acknowledgments

I am grateful to a number of users of the program package for comments and suggestions, although it is impossible to mention their names here. Development of the software has been supported by grants from the Biotechnological and Biological Sciences Research Council (BBSRC) and the Natural Environment Research Councils (NERC).

Literature Cited


Anisimova A, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 24:1219–1228.


Sudhir Kumar, Associate Editor

Accepted April 25, 2007