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_N and d_S) 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, MCMCTREE, 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);
- Estimation of parameters in sophisticated substitution models, including models of variable rates among sites and models for combined analysis of multiple genes (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 (MCMCTREE) 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 1999). 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

Key words: codon models, likelihood, PAML, phylogenetic analysis, software.

E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 24(8):1586–1591. 2007
doi:10.1093/molbev/msm088
Advance Access publication May 4, 2007

Table 1
Example Data Sets Included in the Package to Demonstrate the Analyses

Analysis	File Folder in Examples	Method References	Data Set References
Ancestral reconstruction	stewart.aa	Yang, Kumar, and Nei (1995)	(Stewart, Schilling, and Wilson 1987)
Branch model	lysozyme	Yang (1998)	Messier and Stewart (1997)
Site model	lysin	Yang et al. (2000), Yang and Swanson (2002)	Lee, Ota, and Vacquier (1995)
Branch-site model	lysozyme	Yang and Nielsen (2002)	Messier and Stewart (1997)
Fixed-sites model	lysin	Yang and Swanson (2002)	Lee, Ota, and Vacquier (1995)
ML local-clock models	MouseLemurs	Yoder and Yang (2000), Yang and Yoder (2003)	
Bayesian species divergence dating	DatingSoftBound	Yang and Rannala (2006), Rannala and Yang (2007)	mitochondrial genomes from 7 primates (Cao et al. 1998)

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 (1996), 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 YN00 program implements a number of counting methods for estimating d_N and d_S 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 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 §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 ($\omega = d_N/d_S$) for all nonsynonymous substitutions. This assumption may be unrealistic but simplifies the interpretation of the model. The ω 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 ω ratio to detect positive selection, by calculating d_N and d_S between two sequences, is rarely effective, because the ω 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 ω ratio parameters for different branches on the phylogeny (Yang 1998; Yang

Table 2
Parameters in Site Models

Model	NSsites	p	Parameters
M0 (one ratio)	0	1	ω
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$
M3 (discrete)	3	5	p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 ($p_1 = 1 - p_0$), p, q, $\omega_s > 1$

NOTE.—The site models are implemented using the control variable NSsites in CODEML, and p is the number of free parameters in the ω distribution.

Table 3
Parameters in Branch-Site Model A

Site class	Proportion of sites	Background ω	Foreground ω
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2*a*	$(1 - p_0 - p_1) \cdot p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 \geq 1$
2*b*	$(1 - p_0 - p_1) \cdot p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 \geq 1$

NOTE.—Model A is the alternative hypothesis for the branch-site test of positive selection. The null model fixes $\omega_2 = 1$ for the foreground branch.

and Nielsen 1998). They may be used to detect positive selection acting on particular lineages, without averaging the ω 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 treats the ω ratio for any site (codon) in the gene as a random variable from a statistical distribution, thus allowing ω 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&ω) (Yang et al. 2000). The LRT statistic, or twice the log likelihood difference between the two compared models ($2\Delta\ell$), may be compared against χ^2_2, with critical values to be 5.99 and 9.21 at 5% and 1% significance levels, respectively. When the LRT suggests positive selection, the Bayes empirical Bayes (BEB) method can be used to calculate the posterior probabilities that each codon is from the site class of positive selection under models M2a and M8 (Yang, Wong, and Nielsen 2005). The BEB is an improvement of the early Naïve empirical Bayesian (NEB) method (Nielsen and Yang 1998), and accounts for sampling errors in the ML estimates of parameters in the model.

The *branch-site* models aim to detect positive selection that affects only a few sites on prespecified lineages (Yang and Nielsen 2002). The branches under test for positive selection are called the *foreground* branches, while all other branches on the tree are the *background* branches. The version of the model in common use now is called branch-site model A and is illustrated in table 3 (Yang, Wong, and Nielsen 2005; Zhang, Nielsen, and Yang 2005). In the LRT, branch-site model A is the alternative model, while the simpler null model is model A but with $\omega_2 = 1$ fixed. Because the value $\omega_2 = 1$ is at the boundary of the space of model A, the test statistic $2\Delta\ell$ should be compared with the 50:50 mixture of point mass 0 and χ^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 2005) also suggested the use of χ^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.

The branch-site test requires *a priori* specification of the foreground branches. When multiple branches on the tree are tested for positive selection using the same data set, a correction for multiple testing is required (Anisimova and Yang 2007). A simple and slightly conservative procedure is Bonferroni's correction, which means that the individual test for any branch is considered significant at the level α only if the p-value is $< \alpha/m$, where m is the number of branches being tested using the same data.

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

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

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol. 42:459–468.

Anisimova M, Bielawski JP, Yang Z. 2001. The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. Mol Biol Evol. 18:1585–1592.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 19:950–958.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics. 164:1229–1236.

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.

Bofkin L, Goldman N. 2007. Variation in evolutionary processes at different codon positions. Mol Biol Evol. 24:513–521.

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J Mol Evol. 47:307–322.

Chang BS, Kazmi MA, Sakmar TP. 2002. Synthetic gene technology: applications to ancestral gene reconstruction and structure-function studies of receptors. Methods Enzymol. 343:274–294.

Eyre-Walker A. 1998. Problems with parsimony in sequences of biased base composition. J Mol Evol. 47:686–690.
Felsenstein J. 2005. Phylip: phylogenetic inference program Version 3.6. University of Washington, Seattle.
Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool. 20:406–416.
Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. Syst Biol. 49:652–670.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.
Hartigan JA. 1973. Minimum evolution fits to a given tree. Biometrics. 29:53–65.
Huelsenbeck JP, Bollback JP. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. Syst Biol. 50: 351–366.
Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.
Kimura M. 1977. Prepondence of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 267:275–276.
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.
Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol. 18:352–361.
Koshi JM, Goldstein RA. 1996. Probabilistic reconstruction of ancestral protein sequences. J Mol Evol. 42:313–320.
Kumar S, Tamura K, Nei M. 2005. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.
Lee Y-H, Ota T, Vacquier VD. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol Biol Evol. 12:231–238.
Li W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36:96–99.
Li W-H, Wu C-I, Luo C-C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2:150–174.
Messier W, Stewart C-B. 1997. Episodic adaptive evolution of primate lysozymes. Nature. 385:151–154.
Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. J Mol Evol. 16:23–36.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11:715–724.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.
Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. Heredity. 86:641–647.
Nielsen R, Bustamante C, Clark AG. (13 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.
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.
Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. Syst Biol. 53:47–67.
Pamilo P, Bianchi NO. 1993. Evolution of the *Zfx* and *Zfy* genes - rates and interdependence between the genes. Mol Biol Evol. 10:271–281.
Pupko T, Huchon D, Cao Y, Okada N, Hasegawa M. 2002. Combining multiple data sets in a likelihood analysis: which models are the best? Mol Biol Evol. 19:2294–2307.
Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol. 17:890–896.
Rambaut A, Bromham L. 1998. Estimating divergence dates from molecular sequences. Mol Biol Evol. 15:442–448.
Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Syst Biol.
Ronquist F, Huelsenbeck JP. 2003. MrBayes. 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 19:101–109.
Sawyer SL, Wu LI, Emerman M, Malik HS. 2005. Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci USA. 102:2832–2837.
Self SG, Liang K-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J Am Stat Assoc. 82:605–610.
Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol. 23:7–9.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.
Stewart C-B, Schilling JW, Wilson AC. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature. 330:401–404.
Suchard MA, Kitchen CM, Sinsheimer JS, Weiss RE. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst Biol. 52:649–664.
Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. Mol Biol Evol. 16:1315–1328.
Swofford DL. 2000. PAUP*: phylogenetic analysis by parsimony, version 4. Sinauer Associates. Massachusetts: Sanderland.
Thomas JW, Touchman JW, Blakesley RW, et al. (71 co-authors). 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature. 424:788–793.
Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol. 15:1647–1657.
Thornton J. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nat Rev Genet. 5:366–375.
Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041–1051.
Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol. 42:587–596.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yang Z. 2002. Inference of selection from multiple species alignments. Curr Opinion Genet Devel. 12:688–694.
Yang Z. 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. Acta Zool Sinica. 50:645–656.

Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. Proc Natl Acad Sci USA. 102: 3179–3180.
Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.
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. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409–418.
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, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol. 19:49–57.
Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.
Yang Z, Yoder AD. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol. 52:705–716.
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.
Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol. 17:1081–1090.
Zhang J, Kumar S, Nei M. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol Biol Evol. 14:1335–1338.
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.

Sudhir Kumar, Associate Editor

Accepted April 25, 2007