CABIOS APPLICATIONS NOTE

Vol. 13 no. 5 1997
Pages 555–556

PAML: a program package for phylogenetic analysis by maximum likelihood

Ziheng Yang

Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

Received on March 3 1997; accepted on April 8, 1997

PAML, currently in version 1.2, is a package of programs for phylogenetic analyses of DNA and protein sequences using the method of maximum likelihood (ML). The programs can be used for: (i) maximum likelihood estimation of evolutionary parameters such as branch lengths in a phylogenetic tree, the transition/transversion rate ratio, the shape parameter of the gamma distribution for variable evolutionary rates at sites, and rate parameters for different genes; (ii) likelihood ratio test of hypotheses concerning sequence evolution, such as rate constancy and independence among sites and rate constancy among lineages (the molecular clock); (iii) calculation of substitution rates at sites and reconstruction of ancestral nucleotide or amino acid sequences; and (iv) phylogenetic tree reconstruction by maximum likelihood and Bayesian methods.

The strength of PAML, in comparison with other phylogenetic packages currently available, is its implementation of a variety of evolutionary models. These include several models of variable evolutionary rates among sites, models for combined analyses of multiple gene sequence data and models for amino acid sequences. Multifurcating trees are supported, as well as trees in which some sequences are ancestral to some others. A heuristic tree search algorithm (star decomposition) is used in the package, but tree making is not a strong point of the current version, although work is under way to implement efficient search algorithms. Major programs in the package, as well as the types of analyses they perform, are listed in Table 1. More details are available in the documentation included in the package, written using Microsoft Word.

PAML is distributed free of charge for academic use only. The package, including ANSI C source codes, documentation, example data sets, and control files, can be obtained by anonymous ftp at mw511.biol.berkeley.edu/pub, or from the Indiana molecular biology ftp site at ftp.bio.indiana.edu under the directory Incoming or molbio/evolve. MAC and PowerMac executables are also available, although DOS executables are not prepared yet. Further information about the package is available from the World Wide Web at http://mw511.biol.berkeley.edu/ziheng/paml.html.

Present address: Department of Biology, Galton Laboratory, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK

Table 1. Major programs in the PAML package and their function

Program	Functions
baseml	ML analysis of nucleotide sequences: estimation of tree topology, branch lengths, and substitution parameters under a variety of nucleotide substitution models (JC69, K80, F81, F84, HKY85, TN93, REV); constant or (discrete) gamma rates for sites; auto-discrete gamma model of rate variation and dependence among sites; molecular clock (rate constancy among lineages) or no clock, among-gene and within-gene variation of substitution rates; models for combined analyses of multiple gene data; calculation of substitution rates at sites; reconstruction of ancestral nucleotides
basemlg	ML analysis with a continuous gamma distribution of rates among sites, under a variety of substitution models (JC69, K80, F81, F84 and HKY85)
codonml (codeml with seqtype = 1)	ML analysis of protein-coding DNA sequences using the codon-based model of Goldman and Yang (1994); calculation of the codon-usage table; estimation of synonymous and non-synonymous substitution rates
aaml (codeml with seqtype = 2)	ML analysis of amino acid sequences under several amino acid substitution models (Poisson, Proportional, Dayhoff, Jones *et al.*, [1992], Goldman and Yang, [1994]); constant or gamma rates for sites; molecular clock (rate constancy among lineages) or no clock, among-gene and within-gene variation of substitution rates; models for combined analyses of multiple gene data; calculation of substitution rates at sites; reconstruction of ancestral amino acid sequences
pamp	Parsimony-based analyses for a given tree topology: estimation of the substitution pattern by the method of Yang and Kumar (1996); estimation of the gamma parameter for variable rates across sites by the method of moments, the method of Sullivan *et al.* (1995), and the method of Yang and Kumar (1996); reconstruction of ancestral character states using the algorithm of Hartigan (1973) and an improved parsimony method
mcmctree	Bayesian estimation of phylogenies using DNA sequence data (Rannala and Yang, 1996; Yang and Rannala, 1997). Markov chain Monte Carlo calculation of posterior probabilities of trees
listtree	This program does miscellaneous things, such as listing all rooted and unrooted trees for a given number of species, generating random trees from a birth–death process with species sampling, calculating tree bipartition distances, and simulating nucleotide sequence data sets under a variety of substitution models

Acknowledgements

I thank Nick Goldman, Adrian Friday and Sudhir Kumar for many useful comments on different versions of the program package. I thank Tianlin Wang for the eigen routine used in the package. I also thank a number of users who have reported bugs and made suggestions. Development of the program package has been supported by a grant from the National Science Foundation of China to Z.Y., by NSF and NIH grants to M.Nei, and by an NIH grant to M.Slatkin.

References

Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.

Hartigan,J.A. (1973) Minimum evolution fits to a given tree. *Biometrics*, **29**, 53–65.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Applic. Biosci.*, **8**, 275–282.

Rannala,B. and Yang,Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.*, **43**, 304–311.

Sullivan,J., Holsinger,K.E. and Simon,C. (1995) Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmontine rodents. *Mol. Biol. Evol.*, **12**, 988–1001.

Yang,Z. and Kumar,S. (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.*, **13**, 650–659.

Yang,Z. and Rannala,B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, **14**, 717–724.