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