End-Epi: An application for inferring phylogenetic and population dynamical processes from molecular sequences

Andrew Rambaut, Paul H. Harvey and Sean Nee

Abstract

Motivation. Phylogenetic trees constructed from molecular sequences contain information about the evolutionary or population dynamical processes that created them. Here we describe a computer package (End-Epi) that uses graphical methods to allow researchers to make inferences about these processes from their data. Statistical analyses can be performed to test the consistency of the data with various competing hypotheses.

Availability. End-Epi can be obtained by WWW from http://evolve.zoo.ox.ac.uk/ and by anonymous FTP from ftp://evolve.zoo.ox.ac.uk/packages/End-Epi10.hqx. This file contains the compiled application, the manual and a test tree.

Contact. andrew.rambaut@zoo.ox.ac.uk

Introduction

The branching pattern of phylogenetic trees reconstructed from nucleotide sequences has long been used to make inferences about the evolution of the organisms, or at least the genes, involved. Generally, these trees fall into one of two categories: those whose sequences are sampled from individuals in a population of a single species (within-species trees) and those whose sequences represent different species (between-species trees). Although this application was primarily developed to analyse the former category, researchers may also find it useful for studying between-species trees.

For within-species trees, we have presented methods to infer the population dynamic processes that created them (Nee et al., 1995), based upon the distribution of times between nodes (internode distances). Other methods have been developed using distributions of pairwise sequence differences (Avise et al., 1988; Di Rienzo and Wilson, 1991; Slatkin and Hudson, 1991; Harpending et al., 1993; Marjoram and Donnelly, 1994) and using polymorphisms (Tajima, 1989), but, it has been argued, these methods may be far less efficient than those which use the branching pattern of the entire tree (Felsenstein, 1992). Here we present an application called End-Epi (short for Endemic–Epidemic) which has been developed to extract and graphically display the internode distances from within-species trees and perform statistical tests that allow the researcher to reject some hypotheses about the population dynamic processes involved in favour of others.

The data used in the analyses are the relative time between successive nodes in a phylogenetic tree. Therefore, by necessity, the trees to be analysed will have a time axis with branch lengths proportional to time. Unless independent estimates of the dates of divergence for the sequences exist, from fossil data for example, the trees will have to be constructed under the assumption of a molecular clock. Various programs exist for doing this, the best of which is probably the maximum likelihood approach to be found in Felsenstein’s PHYLIP package (Felsenstein, 1993) in the form of the DNAMLK program. Maximum likelihood is computationally intensive and, for large trees, KITSCH (a distance-based method in the PHYLIP package) may be used. Methods exist to test for the presence of molecular clocks (e.g. Muse and Weir, 1992; Goldman, 1993; Hasegawa et al., 1993; Tajima, 1993), as do more general arguments about their mode and applicability (Kimura, 1987; Gillespie, 1991).

System and methods

This application runs on Apple Macintosh computers and has been tested on the following models: Macintosh SE/30, IIfx, Iici, Quadra and Power Macintosh. It requires at least 1000 kB of memory to run, but for large data sets it may need more. An Apple System 7 operating system (or later) is also required. The speed of the system is not very important as End-Epi is not computationally intensive.

Algorithm

End-Epi will import a tree in a simple text format produced by many phylogeny reconstruction packages such as PHYLIP (Felsenstein, 1993). The tree is displayed in a window from which the user may select all or part of a tree for analysis.

From the tree, or part of it, a graph of the number of lineages through time is produced and displayed in another window. This is constructed by calculating the distances between successive nodes on the tree and plotting the number of lineages cumulatively against them. Initially
the application displays this graph with the lineages axis logarithmically transformed (the semi-log plot).

**Between-species trees**

If the tree represents an entire clade (or a large sample of a clade, say >10%), then this semi-log plot can be used to make inferences about the phylogenetic history of the clade. The analysis of such plots has been explored in detail elsewhere (Nee et al., 1992, 1994a;b; Harvey and Nee, 1993; Harvey et al., 1994; Purvis et al., 1995).

**Within-species trees**

If the tips of the tree represent only a small sample of a large population of sequences (e.g. isolates of a virus, or individuals in a population of animals), then End-Epi can help distinguish between alternative hypotheses about how the population has been changing through time. The derivation and application of these methods have been presented elsewhere (Holmes et al., 1995; Nee et al., 1995). Two natural initial hypotheses are: (i) that the population size has remained constant over the period being studied and (ii) that the population has been increasing exponentially at a constant rate. The initial semi-log plot of lineages-through-time will indicate, by the direction of its curvature, which of these hypotheses should be considered further. We recommend that researchers follow the flowchart shown in Figure 1.

A line which curves up (i.e. the gradient increasing towards the present) is consistent with the constant population size hypothesis (see below); one which curves down is consistent with exponential population growth (see below). We recognize that this may seem counterintuitive, but with a constant size population, pairs of sequences will tend to share a more recent common ancestor than with an exponentially growing population (Nee et al., 1995). If it is not possible to determine a direction of curvature, then a more complex process is required and the analysis is beyond the scope of this application (but see Ong et al., 1996).

**The constant population size hypothesis.** If the semi-log plot curves up, we perform a piece-wise transformation of the time axis by multiplying each inter-node interval by \(n(n - 1)/2\), where \(n\) is the number of lineages at the interval concerned. A plot of cumulative number of lineages against this transformed time will be linear if the tree is consistent with a constant population size. This is called the endemic transform (Nee et al., 1995) and the plot can be displayed in the graph window. If the plot is not linear, then the direction of curvature will indicate whether the population has been increasing or decreasing in size; a steepening towards the present suggests a slowly declining population size and the opposite suggests a slowly increasing population size (see Figure 1).

**The exponential growth hypothesis.** If the population has been growing exponentially at a constant rate, under the epidemic transform defined below, the data will appear to be a linear function of time (with stochastic variation). However, if the population has been growing exponentially at a changing rate, then the graph will appear curved. The direction of curvature indicates whether the rate has been increasing or decreasing (see Figure 1).

The epidemic transform is defined as:

\[
m(t) = \ln \left( \frac{n(0) - n(t)}{n(0)n(t)} \right)
\]