Abstract

Birth-Death (Bi-De) is an application for the Apple Macintosh which simulates the growth of phylogenetic trees using various models of lineage birth and death. The trees produced are intended to be analogous to those reconstructed from molecular sequence data. The user may define a constant birth rate and death rate or a function describing how these rates vary by time or population size. Instantaneous mass extinctions can also be simulated. The package allows the tree produced to be used as a template for the simulated evolution of molecular sequence data under a range of different transition models.

Introduction

Simulation is widely used as a tool for studying molecular evolution because knowledge of the actual phylogeny or the evolutionary processes involved is rare (but see Atchley and Fitch, 1991; Hillis et al., 1992). Simulations of phylogenesis have been used to provide distributions of tree or genetic parameters, e.g. tree balance (e.g. Heard, 1992), branch lengths (Adell and Dopazo, 1994) or pair-wise genetic differences (Slatkin and Hudson, 1991; Harpending et al., 1993) or to test methods of phylogeny reconstruction (Kuhner and Felsenstein, 1994). Such simulations usually fall into two categories: those using a branching process and those using a coalescent process (Hudson, 1990). A branching process starts with a single lineage and progressively grows the tree through time. The coalescent process starts with a collection of contemporary 'twigs' and constructs a tree by successively joining them together into larger branches.

The application we present here, called Bi-De (short for Birth-Death), uses a Markov constant rate birth–death model of lineage splitting and extinction (Raup et al., 1973). Three different processes are simulated. (i) The epidemic process, characterized by a constant per-lineage rate of splitting (births) and a constant per-lineage rate of extinction (deaths). (ii) The endemic process has a constant rate of lineage birth, but for every splitting event, there is a corresponding extinction event. This is the equivalent of the epidemic process with equal birth and death rates, but without any stochastic variation. (iii) The extinction process is an instantaneous event in which a random (phylogenetically independent) sample of the extant lineages at a particular time goes extinct. The terms epidemic, endemic and extinction may not be appropriate for all applications involving the birth–death model but, for convenience, we use them throughout.

Additionally, it is possible for the user to give simple expressions for the birth and death rate as functions of either time or the number of extant lineages. This allows the creation of customized processes, such as density-dependent population dynamics (e.g. the death rate increasing with population size). A number of built-in functions, such as logarithms and exponentials, may be used in the expressions.

When a phylogeny is constructed from molecular sequence data, it consists entirely of branches leading to extant lineages (i.e. the sequences) and contains no information about lineages that have become extinct (see Figure 1). Bi-De simulates the growth of a phylogeny complete with lineage extinction (the actual tree), and then strips out the extinct lineages and the branches leading to them (producing the reconstructed tree).

Using another program, SEQEVOLVE, included in the package, DNA/RNA sequences can be evolved down the reconstructed phylogeny (which can be saved to disk, along with the actual phylogeny). This may be desirable when assessing methods of phylogenetic reconstruction or sequence analysis (Garland et al., 1993; Gaut and Lewis, 1995; Hillis, 1995; Huelsenbeck, 1995).

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 800 kB of memory to run, but for large, complex simulations it will need much more. The Apple System 7 operating system (or later) is also required. The package contains two versions, 'Bi-De FPU' which is for computers with floating-point units and 'Bi-De' which will run on any Macintosh, but is significantly slower.

As well as the compiled Macintosh application, the source code for a less interactive and simplified version is available, called Bi-De95. This is written in ANSI C and will compile...
The simulation consists of a series of periods of evolution, each with a user-specified process under which the phylogeny is generated. Each process is applied in turn and runs until some condition is met. With the exception of the extinction process (which is a single instantaneous event), the periods consist of a series of events at which a lineage can either split or go extinct. The time between events has an exponential distribution with the mean $1/N(\lambda + \mu)$, where $N$ is the number of lineages, and $\lambda$ and $\mu$ are the birth and death rates, respectively. An extant lineage is selected at random and if the event is a birth ($P_{\text{birth}} = \lambda/[\lambda + \mu]$), then this lineage is split, otherwise the lineage is marked extinct. The birth and death rates are either constant for the entire period or vary as a function of $N$ or total time ($t$) using expressions given by the user.

A number of conditions can be specified which, when met, will terminate the current period of evolution. The conditions are: (i) when the phylogeny has a certain number of extant lineages; (ii) when the number of extant lineages changes by a certain amount; and (iii) after a certain amount of time.

For example, the first period could be an epidemic process with a birth rate of 0.2 and a death rate of 0.1, terminating when the number of extant lineages reaches 1000. This could be followed by an extinction event where 80% of the extant lineages go extinct. Finally, a second epidemic period would bring the number of lineages back up to 1000. The output resulting from this example is shown in Figure 2b.

The reconstructed phylogeny

When the simulation is complete, the user can reconstruct the phylogeny from a sample of the extant lineages. This is the equivalent of a tree reconstructed from contemporary sequence data, i.e. it does not contain any of the lineages that went extinct during the course of evolution. The sample can be specified as an exact number or as a percentage of the extant lineages.

The generation of molecular sequence data

Using SEQEVOLVE, sequences may be evolved along the phylogeny produced by Bi-De following a Poisson process with the expected number of substitutions calculated according to a model of molecular evolution. A variety of models of nucleotide substitution are implemented: Jukes and Cantor (1969), Kimura (1980), Felsenstein (1981), Hasegawa et al. (1985), and the DNAML model from PHYLIP (Felsenstein, 1995).

A randomly generated sequence of user-specified length and base composition is used as an ancestral sequence. Alternatively, a specific sequence can be supplied to the program, if desired. The number of repeated simulations can be specified and the resultant sets of sequences are written in PHYLIP format which can be read into most sequence analysis packages.

Implementation

Bi-De is implemented as a standard Macintosh menu-driven application. The main window contains a list of the periods of evolution which can be added to and altered by the user. The parameters of each period in the list can be edited in a dialog box. The list of periods and their parameters may be saved to disk and reloaded later.

When the simulation is run, a phylogeny is generated under the models and parameters of each period in turn. The number of extant lineages (lineages that have not gone extinct by the end of the phylogeny) can be of a potentially realistic order of magnitude ($10^5-10^6$), but this is dependent on the memory available in the computer. A graph window plots the number of lineages in the phylogeny against time as the simulation progresses.
progresses. Once the phylogeny has been created, a number of reconstructions can be created from random samples of extant lineages and these are also plotted (Figure 2).

Both the reconstruction and the actual phylogeny (if not too large) can be saved as machine-readable tree files in Newick format, which is commonly used by phylogeny analysis packages. The lineages-through-time plots can be printed or the data saved as a 'tab-delineated' text file that can be loaded into other graphing or spreadsheet packages.

The non-interactive Bi-De95 reads the parameters from a text file that is created by the user. This version does not allow the user-defined functions for the birth and death rate. However, it will perform replicate simulations saving all the trees to a file and, due to the reduced memory overheads, can potentially produce much larger trees. The tree files produced by either version of Bi-De can be loaded into SEQEVOLVE to create molecular sequence data sets for each simulation.

Discussion

The combination of the phylogeny simulation, Bi-De, and the molecular sequence generator provides a very comprehensive and flexible framework for the exploration of a wide range of molecular evolution and population dynamic processes. The use of the number of lineages plotted through time as a method for inferring evolutionary and population dynamical processes has been described and implemented (Nee et al., 1992, 1995; Harvey and Nee, 1993; Harvey et al., 1994a, b; Nee et al., 1994a, b; Holmes et al., 1995; Purvis et al., 1995). An application for performing these analyses is available from the authors (A. Rambaut, submitted) and the output from the Bi-De can be loaded into this. This allows researchers to explore a range of competing evolutionary hypotheses and compare the results of the simulations with their own data.

Availability

Bi-De can be obtained by WWW from http://evolve.zoo.ox.ac.uk/ and by anonymous FTP from ftp://evolve.zoo.ox.ac.uk/pub/packages/Bi-De10.hqx. This file contains the compiled applications, the source code of the supplementary program, SEQEVOLVE, and the manual. Alternatively, anyone wanting a copy can send a formatted Macintosh disk to the authors.

Acknowledgements

This work was supported by the Wellcome Trust (to A. R. and P. H. H.) and BBSRC (to N. C. G., S. N. and P. H. H.). We would like to thank Eddie Holmes, Andy Purvis and Paul Org for testing, using and commenting on these programs.

References
