CABIOS

Vol. 13 no. 3 1997
Pages 235–238

Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees

Andrew Rambaut[1] and Nicholas C.Grassly

Abstract

Motivation: *Seq-Gen is a program that will simulate the evolution of nucleotide sequences along a phylogeny, using common models of the substitution process. A range of models of molecular evolution are implemented, including the general reversible model. Nucleotide frequencies and other parameters of the model may be given and site-specific rate heterogeneity can also be incorporated in a number of ways. Any number of trees may be read in and the program will produce any number of data sets for each tree. Thus, large sets of replicate simulations can be easily created. This can be used to test phylogenetic hypotheses using the parametric bootstrap.*

Availability: *Seq-Gen can be obtained by WWW from http://evolve.zoo.ox.ac.uk/Seq-Gen/seq-gen.html or by FTP from ftp://evolve.zoo.ox.ac.uk/packages/Seq-Gen/. The package includes the source code, manual and example files. An Apple Macintosh version is available from the same sites.*

Contact: *E-mail: andrew.rambaut@zoo.ox.ac.uk*

Introduction

With few exceptions (e.g. Hillis *et al.*, 1992), the true phylogenetic relationships beween homologous molecular sequences will not be known with certainty. For this reason, simulated nucleotide data sets are widely used to test the efficiency of phylogeny reconstruction methods (e.g. Garland *et al.*, 1993; Tateno *et al.*, 1994; Gaut and Lewis, 1995; Hillis, 1995; Huelsenbeck, 1995).

Simulated DNA sequences can also provide an expectation under a null hypothesis against which real data can be compared. This approach, called the parametric bootstrap, has been used to test the fit of models of nucleotide substitution (Goldman, 1993a), the molecular clock (Adell and Dopazo, 1994; Goldman, 1993b) and competing hypotheses of phylogenetic relationships (Hillis *et al.*, 1996).

The program described here, called Seq-Gen (Sequence Generator), has been designed to be a general purpose simulator that incorporates most of the commonly used (and computationally tractable) models of DNA sequence evolution. The algorithm used is similar to the probability matrix approach described by Schöniger and von Haeseler (1995), but Seq-Gen implements site-specific rate heterogeneity and a wider range of nucleotide subsitution models.

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

[1]*To whom correspondence should be addressed*

The models of substitution

All three models of nucleotide substitution implemented in Seq-Gen are Markov models, and assume that evolution is independent and identical at each site and along each lineage. Almost all models used in the maximum likelihood reconstruction of phylogenies using nucleotide sequences are processes of this type (but see Yang, 1994).

The Hasegawa, Kishino and Yano (HKY) model (Hasegawa *et al.*, 1985) allows for a different rate of transitions and transversions as well as unequal frequencies of the four nucleotides (base frequencies). The parameters required by this model are the transition to transversion ratio (TS/TV) and the base frequencies. There are a number of simpler models that are specific cases of the HKY model (see Figure 1). If the base frequencies are set equal, then the model becomes equivalent to the Kimura 2-parameter (K2P) model (Kimura, 1980). If the TS/TV is set to 0.5 as well, then it becomes equivalent to the Jukes–Cantor (JC69) model (Jukes and Cantor, 1969). If the TS/TV is set to 0.5 and the base frequencies are not equal, then the model is equivalent to the F81 model (Felsenstein, 1981).

The F84 model (Felsenstein and Churchill, 1996), as implemented in DNAML in the PHYLIP package (Felsenstein, 1993), is very similar to HKY, but differs slightly in how it treats transitions and transversions. This model requires the same parameters as HKY.

Finally, the general reversible process (REV) model (e.g.Yang, 1994) allows six different rate parameters and is the most general model that is still consistent with the requirement of being reversible. The six parameters are the relative rates for each type of substitution (i.e. A to C, A to G, A to T, C to G, C to T and G to T). As this is a time-reversible process, the rate parameter of one type of substitution (e.g. A to T) is assumed to be the same as the reverse (e.g. T to A).

Site-specific rate heterogeneity

Site-specific rate heterogeneity allows different sites to evolve at different rates. Two models of rate heterogeneity

$$HKY = \begin{pmatrix} - & \pi_C\beta & \pi_G\alpha & \pi_T\beta \\ \pi_A\beta & - & \pi_G\beta & \pi_T\alpha \\ \pi_A\alpha & \pi_C\beta & - & \pi_T\beta \\ \pi_A\beta & \pi_C\alpha & \pi_G\beta & - \end{pmatrix} \quad F81 = \begin{pmatrix} - & \pi_C\alpha & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & - & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & - & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & \pi_G\alpha & - \end{pmatrix}$$

$$K2P = \begin{pmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{pmatrix} \quad JC69 = \begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix}$$

Fig. 1. The instantaneous rate matrices for the HKY model and the simpler models which are specific cases of it. The nucleotides A, C, G and T are used in alphabetical order, and the top row of a matrix (ignoring the diagonal) is the rate of substitution from A to C, G and T, respectively. The diagonals are the negative sum of the off-diagonal so that each row sums to zero. The symbol π_i is used to denote the frequency of base i, α the rate of transitional substitutions and β the rate of transversional substitutions. Thus, when $\alpha = \beta$, the HKY model becomes the F81 model, but as there are two possible transversions for each transition, the transition–transversion ratio should be set to 1/2. When the base frequencies are equal, HKY becomes K2P and when the transition–transversion ratio is also set to 1/2, it becomes JC69.

are implemented. The first is a codon-based model for simulating protein coding sequences in which the user may specify a different rate for each codon position. For example, more substitutions at the third codon position are synonymous than at the first and second codon positions so this position can be set to evolve at a higher rate.

The second model of rate heterogeneity assigns different rates to different sites according to a gamma distribution (Yang, 1993). The distribution is scaled such that the mean rate for all the sites is 1, but the user must supply a parameter which describes its shape. A low value for this parameter (<1.0) simulates a large degree of site-specific rate heterogeneity and as this value increases the simulated data becomes more rate homogeneous. For reasons of computational tractability, most researchers would reconstruct phylogenies using the discrete gamma model. However, for simulating data, the continuous model is relatively efficient and thus it is used here. For a review of site-specific rate heterogeneity and its implications for phylogenetic analyses, see Yang (1996).

Systems and methods

Seq-Gen is a command-line controlled program written in ANSI C. It should be easily compiled and run on any UNIX system or workstation. The code will also compile on the Apple Macintosh using the Metrowerks Codewarrior compiler. A separate package is available that includes compiled executables, source and instructions for compiling and running the program on these machines. This paper describes the use of Seq-Gen on a UNIX machine. The application requires an amount of memory proportional to the size of each simulated sequence data set. On a Sun Microsystems SPARC 20, it took 11 min to simulate 1000 data sets of 100 sequences and 1000 nucleotides length. The same simulation on an Apple PowerMacintosh 8100/80 took 16 min.

Algorithm

The program simulates a sequence evolving down a branch of the tree using a transition probability matrix (**P**). For a given branch length, this describes the probability (p_{ij}) of a site, which is of state i at one end of the branch, being in state j at the other end. With four states, including the possibility that a site might remain unchanged, this is a four by four matrix with the sum of each row being 1. Taking each base of a sequence in turn, a uniformly distributed random number between 0 and 1 is used to pick one of the four possible outcome states (j) from the row of the transition probability matrix corresponding to the original state (i). The equations for calculating **P** for the different models are developed and discussed elsewhere (F84 model: Thorne *et al.*, 1992; HKY model: Hasegawa *et al.*, 1985; REV model: Yang, 1994). If a model of site-specific rate heterogeneity has been selected, then a different **P** is calculated for each rate. The codon-based rate heterogeneity requires one matrix for each of the three codon positions, whereas the gamma model requires one for each site.

Seq-Gen reads in a tree from a file containing one or more trees in PHYLIP format (Felsenstein, 1993). These trees may be rooted or unrooted, but they must include branch lengths. Each branch length is assumed to denote the mean number of nucleotide substitutions per site that will be simulated along that branch. If this is not the case, then a scaling factor must be supplied that will transform the branch lengths into the mean number of substitutions per site. For example, if the branch lengths are given in units of millions of years (Myr) and the rate of molecular evolution was 0.5 substitutions per site per Myr, then the user should give a scaling factor of 0.5 to obtain the expected amount of substitution along each branch.

With the tree in memory, Seq-Gen creates a random sequence of the specified length and composed of nucleotides at the frequencies given, and this sequence is assigned to the root. If the tree is unrooted, then a root is picked arbitrarily. The program then evolves this sequence down each branch in turn using the appropriate transition probability matrices until it reaches the tips of the tree. The tip sequences are then written out in PHYLIP format. This process is repeated for each replicate requested by the user. If there is more than one tree in the file, then each is read in turn and used to simulate sequences.

Implementation

On a UNIX workstation, Seq-Gen is run by typing its name at the command line. The input file is redirected to the standard input and all the generated data sets are written to

the standard output which may be redirected to a file or the standard input of another program. The switches and parameters that control the program are supplied on the command line. The manual, included with the package, describes how to run the program in detail. An example input file and a sample of the output generated are also included.

Input file format

The input to Seq-Gen is a text file containing one or more trees in the format used by Felsenstein's (1993) PHYLIP package. This format is also used by many other phylogenetics programs such as fastDNAml (Olsen *et al.*, 1994), PAML (Yang, 1995) and MolPhy (Adachi and Hasegawa, 1996). This format is also produced by a program, called Bi-De, that generates trees under a Markov process of lineage birth and death (Rambaut *et al.*, 1996) which, combined with Seq-Gen, allows the simulation of sequences under a wide range of evolutionary hypotheses.

Output file format

The format for the output files was chosen for its simplicity and for the wide range of programs that use it. All of the programs in the PHYLIP package that accept DNA sequences can analyse multiple data sets in the format produced by Seq-Gen. We also include an additional program, called Phy2Nex, which will convert the multiple data set PHYLIP files into NEXUS format for use with the PAUP program (Swofford, 1993).

Discussion

Seq-Gen has been designed to allow the repetitive simulation of DNA sequences under most of the models of nucleotide substitution that are commonly used for the reconstruction of phylogenies. This allows it to be used to generate expected distributions for the testing of nested evolutionary hypotheses (i.e. specific cases of a more general process). This parametric bootstrap approach has been used to test the adequacy of the molecular clock hypothesis (Adell and Dopazo, 1994; Goldman, 1993b) and models of DNA substitution (Goldman, 1993a). A similar approach is used for the testing of alternative hypotheses about the phylogenetic relationships of taxa (Bull *et al.*, 1993; Huelsenbeck *et al.*, 1995; Hillis *et al.*, 1996). Seq-Gen provides a flexible and efficient source of such simulations under a wide range of models of substitution, and may be a useful tool for researchers studying the phylogenetic relationships and molecular evolution of DNA sequences.

Acknowledgements

A.R. was supported by grant 38468 from the Wellcome Trust and N.C.G. by BBSRC. We would like to thank Ziheng Yang for allowing us to use some invaluable code from PAML, and Sophia Kossida for testing, using and commenting on this program.

References

Adachi,J. and Hasegawa,M. (1996) *Programs for Molecular Phylogenetics Based on Maximum Likelihood (MOLPHY), Version 2.3.* The Institute of Statistical Mathematics, Tokyo.

Adell,J.C. and Dopazo,J. (1994) Monte Carlo simulation in phylogenies: An application to test the constancy of evolutionary rates. *J. Mol. Evol.*, **38,** 305–309.

Bull,J.J., Cunningham,C.W., Molineux, I.J., Badgett, M.R. and Hillis, D.M. (1993) Experimental evolution of bacteriophage T7. *Evolution*, **47,** 993–1007.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17,** 368–376.

Felsenstein,J. (1993) *Phylogeny Inference Package (PHYLIP), Version 3.5.* Department of Genetics, University of Washington, Seattle.

Felsenstein,J. and Churchill,G. (1996) A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13,** 93–104.

Garland,T.J., Dickerman,A.W., Janis,C.M. and Jones,J.A. (1993) Phylogenetic analysis of covariance by computer simulation. *Syst. Biol.*, **42,** 265–292.

Gaut,B.S. and Lewis,P.O. (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.*, **12,** 152–162.

Goldman,N. (1993a) Simple diagnostic statistical tests of models for DNA substitution. *J. Mol. Evol.*, **37,** 650–661.

Goldman,N. (1993b) Statistical tests of models of DNA substitution. *J. Mol. Evol.*, **36,** 182–198.

Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22,** 160–174.

Hillis,D.M. (1995) Approaches for assessing phylogenetic accuracy. *Syst. Biol.*, **44,** 3–16.

Hillis,D.M., Bull,J.J., White,M.E., Badgett,M.R. and Molineux,I.J. (1992) Experimental phylogenetics: Generation of a known phylogeny. *Science*, **255,** 589–592.

Hillis,D.M., Mable,B.K. and Moritz,C. (1996) Applications of molecular systematics: the state of the field and a look to the future. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds.), *Molecular Systematics*, Sinauer Associates, Inc., Sunderland.

Huelsenbeck,J.P. (1995) Performance of phylogenetic methods in simulation. *Syst. Biol.*, **44,** 17–48.

Huelsenbeck,J.P., Hillis,D.M. and Jones,R. (1995) Parametric bootstrapping in molecular phylogenetics: Applications and performance. In Ferraris,J. and Palumbi,S. (eds.), *Molecular Zoology: Strategies and Protocols*, Wiley, New York.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.) *Mammalian Protein Metabolism*, Academic Press, New York, pp. 21–123.

Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16,** 111–120.

Olsen,G.J., Matsuda,H., Hagstrom,R. and Overbeek,R. (1994) fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Applic. Biosci.*, **10,** 41–48.

Rambaut,A., Grassly,N.C., Nee,S. and Harvey,P.H. (1996) Bi-De: An application for simulating phylogenetic processes. *Comput. Applic. Biosci.*, in press.

Schöniger,M. and von Haeseler,A. (1995) Simulating efficiently the evolution of DNA sequences. *Comput. Applic. Biosci.*, **11,** 111–115.

Swofford,D.L. (1993) *Phylogenetic analysis using parsimony (PAUP), Version 3.1.1.* Illinois Natural History Survey, Champaign.

Tateno,Y., Takezaki,N. and Nei,M. (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony method when substitution rate varies with site. *Mol. Biol. Evol.*, **11,** 261–277.

Thorne,J.L., Kishino,H. and Felsenstein,J. (1992) Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.*, **34,** 3–16.

Yang,Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, **10,** 1396–1401.
Yang,Z. (1994) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, **39,** 105–111.
Yang,Z. (1995) *Phylogenetic Analysis by Maximum Likelihood (PAML), Version 1.1.* Institute of Molecular Evolutionary Genetics, The Pennsylvania State University.
Yang,Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Tr. Ecol. Evol.*, **11,** 367–372.

Received on August 22, 1996; revised and accepted on October 30, 1996