MySSP: Non-stationary evolutionary sequence simulation, including indels

Michael S. Rosenberg

Center for Evolutionary Functional Genomics and the School of Life Sciences, Arizona State University, Tempe, AZ, USA

Abstract: MySSP is a new program for the simulation of DNA sequence evolution across a phylogenetic tree. Although many programs are available for sequence simulation, MySSP is unique in its inclusion of indels, flexibility in allowing for non-stationary patterns, and output of ancestral sequences. Some of these features can individually be found in existing programs, but have not all have been previously available in a single package.

Keywords: Sequence Simulation, DNA, Indels, Non-stationarity

Introduction

Simulation of molecular sequence evolution has become a fundamental part of comparative genomic and bioin-formatics analysis. Simulation has proven particularly useful for testing the efficacy of bioinformatics methods and techniques under a variety of conditions and assumptions (or violations thereof), including, for example, phylogenetic analysis (Hillis 1995; Nei 1996; Takahashi and Nei 2000; Rosenberg and Kumar 2003; Huelsen-beck and Rannala 2004, just to name a few) and sequence alignment (Keightley and Johnson 2004; Pollard et al 2004; Rosenberg 2005). Many programs are available for simulating molecular sequence evolution, including Evolver (PAML) (Yang 1997), Seq-Gen (Rambaut and Grassly 1997), ROSE (Stoye et al 1998), and DAWG (Cartwright 2005), each with its own set of strengths and weaknesses. The program presented here, MySSP, has been gradually developed over a series of projects (including, eg, Rosenberg and Kumar 2001; Rosenberg and Kumar 2003; Gadagkar et al 2005; Rosenberg 2005) and is being made publicly available because of some unique features, individually and in combination, which are not found in other available packages.

As with many similar programs, given a fixed tree (supplied by the user) MySSP constructs an initial DNA sequence at the root of the tree and simulates evolution across the tree using a variety of common models of DNA evolution, including Jukes-Cantor (Jukes and Cantor 1969), Kimura two-parameter (Kimura 1980), equal input, Hasegawa-Kishino-Yano (Hasegawa et al 1985), and the general time-reversible model. Rate variation among sites can optionally be modeled with the standard gamma-distribution for any of these models. Multiple genes with different parameters and models can be simulated simultaneously. MySSP is designed for large-scale studies, including simulation of multiple replicates and outputs sequences into NEXUS, MEGA, or FASTA formats. MySSP has a fairly simple GUI for basic use, but also has a specialized batch script interpreter to allow for more complicated or large-scale simulations.

Where MySSP becomes unique relative to most other simulation programs is (1) its ability to simulate insertion and deletion events; (2) its ability to allow simulation of nonstationary processes and models across the tree; and (3) its option to output ancestral sequences. Two of these features (1 and 3) can individually be found in existing programs, but not all have been previously available in a single package. Each is described in turn.

Simulation of Insertions and Deletions

Insertions and deletions (indels) are a common component of sequent evolution, but historically have not been included in most simulation packages; only two are known to include indel evolution: ROSE (Stoye et al 1998) and DAWG (Cartwright 2005). MySSP simulates insertions and deletions using simple Poisson models for rate and size distribution of insertion and deletion events (modeled separately, parameters provided by the user). One advantage of MySSP is that the output sequences are aligned

Correspondence: Michael S Rosenberg, School of Life Sciences, PO Box 874501, Tempe, AZ 85287-4501, Phone: 480-965-1578, Fax: 480-649-6899, E-mail: msr@asu.edu.

correctly, ie, the output sequences include gaps such that aligned sites across sequences represent true homologies. This gives one a baseline "true alignment" that can be used to contrast with the results from removing the gaps from the output sequences (a trivial exercise) and running them through a standard alignment program.

Non-stationary processes and models

A common concern in molecular sequence analysis is whether the evolutionary process is stationary across a tree. While there are many possible models of sequence evolution, the majority of simulation programs assume that whatever model is specified is constant throughout the tree. MySSP allows the user to change the evolutionary model for each and every branch, if they desire. One can completely change every aspect of the model, including basic substitution pattern (JC, HKY, etc.), transition-transversion bias, gamma distributed rate variation, equilibrium nucleotide frequencies, and indel rate and size. One can also change the basic rate of substitution for a branch, increasing or decreasing it relative to that found on the model tree. This flexibility allows one to much more easily examine the effects of non-stationary processes on bioinformatics analysis, eg, using a single "average" model in maximum likelihood phylogenetic analysis. The ability to completely change the model for each and every aspect of the tree is unique among simulation programs.

Ancestral sequences

MySSP also includes an option for outputting ancestral sequences, that is, the sequence found at each and every node on the tree. This may be useful for those wishing to test methods of ancestral state reconstruction or for whom tracing changes from ancestral sequences may be important. Ancestral sequence output is available from Evolver (Yang 1997) and Seq-Gen (Rambaut and Grassly 1997), but not in combination with indel and non-stationary simulation.

Availability

The program and documentation can be freely downloaded from http://lsweb.la.asu.edu/rosenberg. It runs natively under all 32-bit Windows operating systems and has also successfully been used under Linux emulators. Source code is available on request.

Acknowledgements

Thanks to S. Kumar, S. Gadagkar, T. H. Ogden, and anonymous reviewers for advice and suggestions on the development of the program. This work is partially supported by NIH R03-LM008637 and Arizona State University.

References

Cartwright RA. 2005. DAWG: DNA Assembly with Gaps. http://scit.us/dawg.

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. Journal of Experimental Zoology B Molecular and Devel-opmental Evolution, 304B:64-74.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol, 22:160-74.

Hillis DM. 1995. Approaches for assessing phylogenetic accuracy. Syst Biol, 44:3-16.

Huelsenbeck JP, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst Biol, 53:904-13.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In Munro HN, ed. Mammalian Protein Metabolism. New York: Academic Press. p 21-132.

Keightley PD, Johnson T. 2004. MCALIGN: Stochastic alignment of non-coding DNA sequences based on an evolutionary model of sequence evolution. Genome Res, 14:442-50.

Kimura M. 1980. A simple method for estimating evolutionary rates of base subsitutions through comparative studies of nucleotide sequences. J Mol Evol, 16:111-20.

Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Ann Rev Gen, 30:371-403.

Pollard DA, Bergman CM, Stoye J et al. 2004. Benchmarking tools for the alignment of functional noncoding DNA. BMC Bioinformatics, 5:6.

Rambaut A, Grassly NC. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in Bioscience, 13:235-8.

Rosenberg MS. 2005. Evolutionary distance estimation and fidelity of pair wise sequence alignment. BMC Bioinformatics, 6:102.

Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. PNAS, 98:10751-6.

Rosenberg MS, Kumar S. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol Biol Evol, 20:610-21.

Stoye J, Evers D, Meyer F. 1998. Rose: Generating sequence families. Bioinformatics, 14:157-63.

Takahashi K, Nei M. 2000. Efficiencies of fast algorthims of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol, 17:1251-8.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Computer Applications in Bioscience, 13:555-6.