BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 18 2006, pages 2313–2314
doi:10.1093/bioinformatics/btl387

Phylogenetics

Serial NetEvolve: a flexible utility for generating serially-sampled sequences along a tree or recombinant network

Patricia Buendia and Giri Narasimhan*

[1]Bioinformatics Research Group (BioRG), School of Computing and Information Science, Florida International University, Miami, FL 33199, USA

Received on May 1, 2006; revised on July 5, 2006; accepted on July 6, 2006
Advance Access publication July 14, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Summary: *Serial NetEvolve* is a flexible simulation program that generates DNA sequences evolved along a tree or recombinant network. It offers a user-friendly Windows graphical interface and a Windows or Linux simulator with a diverse selection of parameters to control the evolutionary model. *Serial NetEvolve* is a modification of the *Treevolve* program with the following additional features: simulation of serially-sampled data, the choice of either a clock-like or a variable rate model of sequence evolution, sampling from the internal nodes and the output of the randomly generated tree or network in our newly proposed NeTwick format.

Availability: From website http://biorg.cis.fiu.edu/SNE

Contacts: giri@cis.fiu.edu

Supplementary information: Manual and examples available from http://biorg.cis.fiu.edu/SNE

1 INTRODUCTION

There has been considerable recent interest in understanding recombination and developing new methods for recombination detection and phylogenetic network reconstruction (Martin *et al*., 2005; Nakhleh *et al*., 2003). Another area that has received increased attention is the analysis of serially-sampled sequence data derived from viruses sampled from a single infected patient (Drummond and Rodrigo, 2000; Nickle *et al*., 2003). The simulation program, *SeqGen* (Rambaut and Grassly, 1997), assists in the evaluation of phylogenetic software by generating synthetic sequences evolved along a user specified tree. For a specified set of parameters, *Treevolve* (Grassly *et al*., 1999) generates a coalescent tree (Kingman, 1982) or recombinant network (Hudson, 1983) and evolves a set of sequences along that structure. Neither supports serial samples. Moreover, although *Treevolve* does not require a user specified topology, it does not output the topology it generates. Here we present *Serial NetEvolve*, a modification of *Treevolve*, which generates serially-sampled sequences along a randomly generated reticulate network. It also provides the network topology, which may be used in recombination detection programs. The option to generate a recombinant network is not featured in the recently published *Serial SIMCOAL* (Anderson *et al*., 2005) and in the earlier *Serial Coalescent Simulator* (Drummond and Strimmer, 2001). *Serial NetEvolve* offers a flexible set of population parameters (migration rate, population growth rate, among others) present in the original *Treevolve* and additional options for the generation of both trees and recombinant networks.

*To whom correspondence should be addressed.

2 PROGRAM FEATURES

The new features of *Serial NetEvolve* are described in detail below. Additional details on how to use the features of *Serial NetEvolve* may be found at the software website mentioned above. The program GUI was implemented in Visual Basic 6 and the simulator was written in ANSI C. Source code and executables are available for Windows 2000/XP and Linux.

2.1 Serially-sampled sequences

A theoretical framework for generating serial samples in the coalescent was developed (Rodrigo and Felsenstein, 1999) and implemented in Java in *Serial Coalescent Simulator* (Drummond and Strimmer, 2001). In *Serial NetEvolve*, we implemented the technique by modifying *Treevolve* to have sequences assigned to different time points instead of assigning them to the zero-time baseline.

2.2 Molecular clock

Distance to the root and time are equivalent when enforcing a molecular clock. To simulate datasets with a constant rate of evolution (enforced molecular clock), sequences from the same sampling time were assigned to the same distance from the root, as implemented in *Serial Coalescent Simulator*. For the variable rates setting, all sequences (independent of the sampling time they belonged to) were assigned to randomly generated distances from the root. For consistency, distances of the nodes were constrained to decrease as one traverses towards the root.

2.3 Internal node sampling

Serial NetEvolve allows the user to set the probability of internal node sampling as a parameter. Sampling of internal nodes makes sense only if the effective population is small (Drummond and Rodrigo, 2000), which in turn is more probable when the length of the sequences sampled is small. It is worth noting that fairly short sequences [$\sim$700 bp in Shankarappa *et al*. (1999)] were used in many large studies of viral populations, and that effective population size for serially-sampled HIV-1 was estimated to be only 4232.2 and negatively correlated to the evolutionary rate (Seo *et al*., 2002).

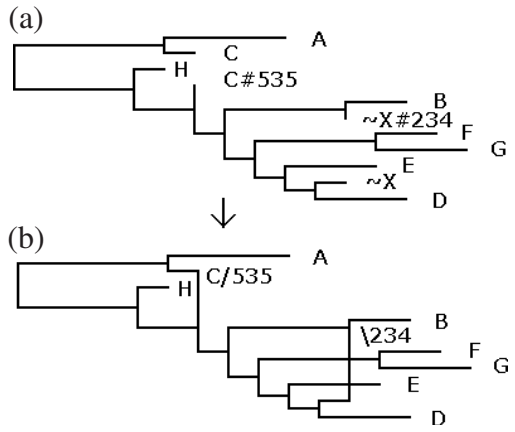

Fig. 1. The (**a**) tree representation and the (**b**) network representation for a network represented by the NeTwick code: ((A:4,C:1):5,((((B:2,~X#234:0):4,((E:3,(~X:1,D:3):1):1,(F:2,G:3):4):1):1,C#535:0):2,H:1):4).

2.4 Tree or network output

Treevolve was modified to output the coalescent tree or network. When the recombination rate is zero the tree is output to a file in Newick format, which represents trees using nested parentheses (Felsenstein 1999, http://evolution.genetics.washington.edu/phylip/newicktree.html). If internal nodes are sampled, they are assigned to zero-length branches. In order to write a recombinant network to a file, we devised the 'NeTwick' format, a variant of the Newick format, incorporating additional information (breakpoint position, right and left parent) to represent recombinant nodes. Unlike tree nodes, recombinant nodes have more than one parental node. In *Serial NetEvolve*, we (arbitrarily) chose the left parental node of a recombinant sequence to appear twice in the NeTwick format to indicate the linking relationship. One of the copies of the left parental node appears followed by the symbol '#', along with the breakpoint position and it represents a link, not a taxon. If the left parent was not sampled, it also appears with a '∼' prefix. Figure 1 shows a network with nine taxa and its tree equivalent. The left parent X was not sampled (indicated by the ∼), but is present in the tree to indicate the linking relationship as shown in the network. In the proposed network representation a backward (forward, respectively) slash followed by the breakpoint number indicates whether the left parent is below (above, respectively) in a horizontally drawn network. The advantage of making two copies of the left parental node of every recombinant node is that the network can then be represented by an equivalent tree. The tree can be viewed using any tree-viewing program; a network viewer is currently being developed. NeTwick supports tree and network polytomies, but is restricted to one child per recombinant parent.

3 DISCUSSION AND CONCLUSION

The power of *Serial NetEvolve* lies in the ease with which it is possible to generate a collection of datasets using a wide range of parameters with the goal of comparing different programs that analyze any aspect of serially-sampled sequence data. For instance, it is possible to compare with relative ease the topologies of the output of several programs. An example of such an experiment can be found in the supplemental website. *Serial NetEvolve* differs from the majority of simulation methods in that it incorporates both serial sampling and recombination along with additional features (heterogeneous evolution rate, sampling of internal nodes), while at the same time, maintaining the population parameters from *Treevolve* (migration rate, population growth rate, among others). Analysis of serially-sampled data has applications in the study of the evolution of fast-evolving DNA and RNA viruses, for which it is often the case that recombination is an integral part of their life cycle. The recombination rate in HIV-1 is one of the highest among all organisms (Rambaut *et al*., 2004). Thus coalescent simulators for serial samples need to incorporate a recombination feature in their sequence generation, as is the case with *Serial NetEvolve*. Finally, the option to output the tree or network to a file is critical for automated evaluations of the resulting tree or network. Note that network topologies can now be compared using a measure proposed by Nakhleh *et al*. (2003), who extended the well-known Robinson–Foulds (RF) score (Robinson and Foulds, 1981) for tree topologies.

ACKNOWLEDGEMENTS

P.B. was supported by MBRS-RISE Fellowship (NIH/NIGMS R25GM61347). G.N. was supported in part by NIH Grant P01 DA15027-01.

Conflict of Interest: none declared.

REFERENCES

Anderson,C.N.K. *et al*. (2005) Serial SIMCOAL: a population genetics model for data from multiple populations and points in time. *Bioinformatics*, **21**, 1733–1734.

Drummond,A. and Rodrigo,A.G. (2000) Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA (sUPGMA). *Mol. Biol. Evol*., **17**, 1807–1815.

Drummond,A. and Strimmer,K. (2001) PAL: An object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, **17**, 662–663.

Felsenstein,J. (1999) The Newick tree format.

Grassly,N. *et al*. (1999) Population dynamics of HIV-1 inferred from gene sequences. *Genetics*, **151**, 427–438.

Hudson,R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol*., **23**, 183–201.

Kingman,J.F.C. (1982) The coalescent. *Stochastic Process. Appl*., **13**, 235–248.

Martin,D.P. *et al*. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**, 260–262.

Nakhleh,L. *et al*. (2003) Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. *Pac. Symp. Biocomput*., 315–326.

Nickle,D.C. *et al*. (2003) Evolutionary indicators of Human Immunodeficiency Virus type 1 reservoirs and compartments. *J. Virol*., **77**, 5540–5546.

Rambaut,A. and Grassly,N. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci*., **13**, 235–238.

Rambaut,A. *et al*. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet*., **5**, 52–61.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci*., **53**, 131–147.

Rodrigo,A. and Felsenstein,J. (1999) Coalescent approaches to HIV-1 population genetics. In Crandall,K. (ed.), *Molecular Evolution of HIV*. Johns Hopkins University Press, MD, pp. 233–272.

Seo,T.-K. *et al*. (2002) Estimation of effective population size of HIV-1 within a host: A pseudomaximum-likelihood approach. *Genetics*, 1283–1293.

Shankarappa,R. *et al*. (1999) Consistent viral evolutionary changes associated with the progression of HIV 1 infection. *J. Virol*., **73**, 10489–10502.