BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 15 2006, pages 1921–1923
doi:10.1093/bioinformatics/btl191

Phylogenetics

NetGen: generating phylogenetic networks with diploid hybrids

M. M. Morin* and B. M. E. Moret
Department of Computer Science, University of New Mexico, Albuquerque, NM, USA

Received on April 8, 2006; revised on May 13, 2006; accepted on May 14, 2006
Advance Access publication May 22, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Summary: NetGen is an event-driven simulator that creates phylogenetic networks by extending the birth–death model to include diploid hybridizations. DNA sequences are evolved in conjunction with the topology, enabling hybridization decisions to be based on contemporary evolutionary distances. NetGen supports variable rate lineages, root sequence specification, outgroup generation and many other options. This note describes the NetGen application and proposes an extension of the Newick format to accommodate phylogenetic networks.

Availability: NetGen is written in C and is available in source form at http://www.phylo.unm.edu/~morin/

Contact: morin@cs.unm.edu

Phylogenetic simulation and reconstruction methods usually assume a tree topology with independent events of speciation and extinction. However, reticulate events such as hybridization and lateral gene transfer are known to occur along some evolutionary paths at the inter-species level. The presence of such events transforms a tree into an evolutionary *network*. Over the last two decades, phylogenetic networks have received increased attention at both the inter-species (e.g. Gusfield, 2005; Moret *et al*., 2004; Nakhleh *et al*., 2003) and the intra-species levels (e.g. Bandelt *et al*., 1999; Posada and Crandall, 2001; Excoffier and Smouse, 1994). Software aimed at inter-species networks is limited [e.g. SplitsTree (Huson and Bryant, 2006) and T-REX (Makarenkov, 2001)] while there are more packages available at the intra-species level [e.g. Treevolve (Grassly *et al*., 1999), Network (FluxusEngineering, 1999, http://www.fluxus-engineering.com/), TCS (Clement *et al*., 2000), Arlequin (Schneider *et al*., 2000), UMP (Cassens *et al*., 2005)]. However, a simulator at the inter-species level to generate evolutionary networks does not exist, a lack that hampers the assessment and further development of network reconstruction methods.

NetGen (Morin, 2006) is an event-driven simulator that extends the birth–death model (Renshaw, 1991) to include diploid hybrids. Hybridization decisions are made according to the DNA sequences of contemporary lineages; thus, in contrast to tree-based simulators, sequences co-evolve with the topology. Options include variable rate lineages, hybridization limits, choice of root sequences, outgroup generation and deviation from ultrametricity. Written in C, NetGen is command-line driven with text files for input and output. It is composed of three modules. The first module, NG, contains the simulation logic; the second provides the interface between NG and the third, which is a sequence generator. We have chosen the popular Seq-Gen (Rambaut and Grassly, 1977) as our sequence generator, but any desired generator can be used with modification to the second module. Robust simulation studies can be conducted as NetGen utilizes individual process identifiers to distinguish the communication files for separate simulations. Furthermore, a user-specified seed for a pseudorandom number generator ensures reproducibility of runs.

A priority queue tracks all events. Currently three types of events are implemented: birth, death and diploid hybridization. The queue structure enables the addition of other types of events such as lateral gene transfer, polyploid hybridizations and mass extinction. Events are scheduled according to exponential interarrival times. At the creation of a new lineage, event times are generated for each of the event types using the corresponding user-specified rates and the earliest of these generated events is retained and placed on the queue.

The network generation process starts with two active lineages, each with a future event scheduled on the queue. The simulation advances by processing the next queued event. Once completed, events are logged and removed from the queue. This process continues until either the desired number of extant taxa is reached or the event queue is empty. In the former, normal case, we choose the final end time randomly between the last completed event and the first remaining event on the queue (if any); this choice avoids artificially short branch lengths and matches reality well, since the sampling of modern taxa occurs at what amounts to a randomly chosen time between two evolutionary events.

NetGen tracks both evolutionary and clock-based branch lengths. When requesting a sequence, the evolutionary branch length is passed to the sequence generator so the appropriate amount of evolutionary change can be simulated. The clock time is required to order the queue and to identify all active lineages when searching for a potential second parent in a hybridization event.

The network aspect of NetGen comes from the implementation of diploid hybridization. As described above, a lineage is assigned a future event when it is created. If the scheduled event is a hybridization and the maximum number of allowed hybrids has not been reached, a second contemporary lineage is sought to form a pair of parent lineages for the hybrid. First branch lengths and sequences are updated for all active lineages. A second parent for the hybridization event is then chosen according to the user-specified method of (1) minimum hamming distance, (2) minimum evolutionary distance or (3) random choice. In the first two cases, a user has the option to additionally specify a threshold above which a hybridization will not be performed. Each parent randomly contributes half of its chromatids in each chromosome to create a new species, while propagating its original lineage (Fig. 1).

*To whom correspondence should be addressed.

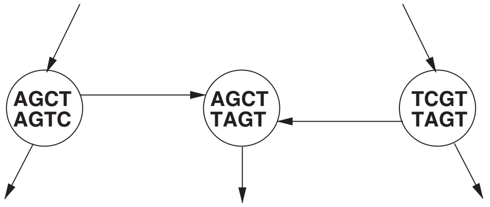

Fig. 1. Graphical representation of a diploid hybrid event. Each parent donates half of its chromatid sequences to the hybrid species.

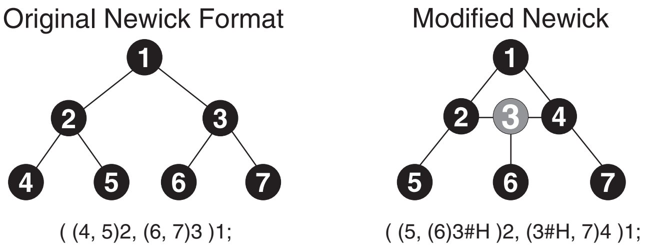

Fig. 2. Regular versus modified Newick formats. Phylogeny on the right includes a hybrid node (3) denoted by *#H*.

Table 1. Predicted and observed mean population growth rates

Rates B	 D	 H	Expected growth $B - D + H$	Measured growth fitted slope ± σ
2.0	—	—	2.0	2.004 ± 0.013
2.0	0.5	—	1.5	1.498 ± 0.020
—	—	1.0	1.0	0.998 ± 0.008
2.0	—	1.0	3.0	2.999 ± 0.027
2.0	0.5	1.0	2.5	2.492 ± 0.031

Ten iterations per scenario, each with 25 000 extant taxa.

The continuing lineage of the first parent and the hybridized lineage are assigned their future events based on the specified event rates. The lineage of the second parent already had an event queued prior to the hybridization and so proceeds without requiring the generation of a new event.

NetGen allows the user to declare the number of chromosomes and chromatids (memory permitting) as well as the root sequence(s). If root sequence information is not provided, it is created by the sequence generator. Outgroups can be generated according to user-specified (dis)similarity bounds. NetGen creates by default ultrametric networks (with respect to evolutionary branch lengths) and fixed event rates across all lineages, as is typical in the phylogenetic community, but it also allows for the exploration of different model behaviors. If non-ultrametric networks are desired, the evolutionary branch lengths are deviated according to a gamma distribution. For the variable-rate lineage option, each lineage is assigned positive event rates from normal distributions specified by the user.

Under the birth–death model, the mean population size is described by $n = n_0 e^{(B-D)\cdot t}$, where n_0 is the initial population size, t is the time and B and D are the birth and death rates, respectively (Renshaw, 1991). Like a birth, hybridization adds one new lineage to the phylogeny; therefore the new mean population size is now $n = n_0 e^{(B-D+H)\cdot t}$, where H is the hybridization rate. Thus the natural log of the population size grows linearly in time with a slope of $B - D + H$—something we verified experimentally for 25 000 extant taxa and a variety of values for B, D and H (Table 1).

With phylogenetic network visualization in mind, NetGen includes a modified version of the Newick tree format which we have developed to describe networks. Figure 2 shows a tree and a network, with corresponding Newick strings. Node 3 of the network is a hybrid node, denoted by the addition of `#H`. Lateral gene transfer, another reticulate evolutionary event, can be represented in this format as well by annotating such nodes with `#LGT`. Generating this format requires knowledge about the node type and which nodes have been visited, so as to avoid traversing their subtrees more than once. This format is a depth-first traversal and reverts to Newick tree format when network nodes are not present.

NetGen is a novel simulator for creating phylogenetic networks with diploid hybrids. DNA sequences are evolved with the topology allowing the hybridization decisions to be based on sequence similarity. In combination with the tripartition measure of Moret *et al.* (2004), NetGen enables us to assess the performance of network reconstruction tools. Planned enhancements include adding other events such as lateral gene transfer, providing more complex models of birth [e.g. inheritability of speciation rates (Heard, 1996)], incorporating the tripartition measure of Moret *et al.* (2004) and packaging the whole as a module for Mesquite (Maddison and Maddison, 2001, mesquiteproject.org).

ACKNOWLEDGEMENTS

This work is funded by the National Science Foundation under grants IIS 01-21377, DEB 01-20709 and EF 03-31654.

Conflict of Interest: none declared.

REFERENCES

Bandelt,H. *et al.* (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, **16**, 37–48.

Cassens,I. *et al.* (2005) Evaluating intraspecific ‘Network’ construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? *Syst. Biol.*, **54**, 363–372.

Clement,M. *et al.* (2000) TCS: a computer program to estimate gene genealogies. *Mol. Ecol.*, **9**, 1657–1660.

Excoffier,L. and Smouse,P. (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics*, **136**, 343–359.

FluxusEngineering. (1999) `Network` software package.

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

Gusfield,D. (2005) Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. *J. Comput. Syst. Sci.*, **70**, 381–398.

Heard,S. (1996) Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution*, **50**, 2141–2148.

Huson,D. and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **2**, 254–267.

Maddison,W. and Maddison,D. (2001) Mesquite: a modular system for evolutionary analyses, version 0.98.

Makarenkov,V. (2001) T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, **17**, 664–668.

Moret,B. *et al.* (2004) Phylogenetic networks: modeling, reconstructability, and accuracy. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **1**, 13–23.

Morin,M. (2006) NetGen Release Notes, Epsilon Version (Phylogenetic Network Generation Application). *Technical Report TR-CS-2006–05,* University New Mexico, Albuquerque, NM.

Nakhleh,L., Sun,J., Warnow,T., Linder,R., Moret,B. and Tholse,A. (2003) Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB'03),* Hawaii. World Scientific Publishing, Hackensack, NJ, pp. 315–326.

Posada,D. and Crandall,K. (2001) Intraspecific gene geneaologies: trees grafting into networks. *Trends Ecol. Evol.*, **16**, 37–45.

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.

Renshaw,E. (1991) *Modelling Biological Populations in Space and Time*. Cambridge University Press, New York, NY.

Schneider,S., Roessli,D. and Excoffier,L. (2000) Arlequin: a software for population genetics data analysis. *Technical report, ver 2.000*. Genetics and Biometry Lab, Department of Anthropology, University of Geneva, Geneva, Switzerland.