# Characterization of Reticulate Networks Based on the Coalescent with Recombination

*Miguel Arenas,* Gabriel Valiente,† and David Posada**

*Department of Biochemistry, Genetics and Immunology, University of Vigo, E-36310 Vigo, Spain; and †Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain

Phylogenetic networks aim to represent the evolutionary history of taxa. Within these, reticulate networks are explicitly able to accommodate evolutionary events like recombination, hybridization, or lateral gene transfer. Although several metrics exist to compare phylogenetic networks, they make several assumptions regarding the nature of the networks that are not likely to be fulfilled by the evolutionary process. In order to characterize the potential disagreement between the algorithms and the biology, we have used the coalescent with recombination to build the type of networks produced by reticulate evolution and classified them as regular, tree sibling, tree child, or galled trees. We show that, as expected, the complexity of these reticulate networks is a function of the population recombination rate. At small recombination rates, most of the networks produced are already more complex than regular or tree sibling networks, whereas with moderate and large recombination rates, no network fit into any of the standard classes. We conclude that new metrics still need to be devised in order to properly compare two phylogenetic networks that have arisen from reticulating evolutionary process.

## Introduction

Phylogenetic networks represent the evolutionary relationships of taxa, including sequences, genes, chromosomes, genomes, or species. There are different types of phylogenetic networks, but here, we are interested in reticulate networks, which provide an explicit representation of evolutionary history, meaning that internal "nodes" represent ancestral species, and nodes with more than two parents correspond to reticulate events such as recombination, hybridization or lateral gene transfer (Huson and Bryant 2006). Reticulate networks have been extensively used in evolutionary studies, especially at the population level, where reticulate events are, in general, quite common (Posada and Crandall 2001).

Very few studies have tried to assess the performance of the algorithms used to reconstruct phylogenetic networks (Cassens et al. 2003, 2005; Jin et al. 2007), and only recently, a comprehensive computer simulation study has been completed (Woolley et al. 2008). One of the problems that arose during these studies was the comparison of reticulate networks. Although several comparison metrics have been already introduced in the literature (Baroni et al. 2004; Moret et al. 2004; Cardona, Llabrés, et al. 2008a, 2008b; Cardona, Rosselló, and Valiente 2008b, 2008c; Cardona G, Llabrés M, Rosselló F, Valiente G, Metrics for phylogenetic networks II: Nodal and triplets metrics, unpublished data; Cardona G, Llabrés M, Rosselló F, Valiente G, Recent advances in metrics for phylogenetic networks, unpublished date), all of them have specific requirements regarding the nature of the networks to be compared in order to be "perfect," that is, to have the metric properties of nonnegativity, separation, symmetry, and triangle inequality.

Before continuing, it will be necessary to give some definitions. A network contains nodes (vertices) and "branches" (edges) that connect them. In general, we will refer to "rooted networks," with a direction from the past to the present (i.e., directed graphs) that allows for the identification of "parent" and "child" (descendant) nodes. "External nodes" (leaves) have no children, whereas "internal nodes" have two. The "root node" is the oldest node and has no parents. "Tree nodes" have just one parent, whereas "hybrid nodes" have two parents. When two nodes share the same parent they are "siblings." Depending on the relationships among the nodes that occur in a network, these can be classified as "tree sibling" (Cardona, Llabrés, et al. 2008a; Cardona, Rosselló, and Valiente 2008a), where every hybrid node has at least one sibling that is a tree node; "tree child" (Cardona, Llabrés, et al. 2008b; Cardona, Rosselló, and Valiente 2008a, 2008b, 2008c; Cardona G, Llabrés M, Rosselló F, Valiente G, unpublished data), in which every internal node has at least one child that is a tree node; "regular" (Baroni et al. 2004), where the set of descendant leaves (clusters) of the nodes are all distinct; "galled trees" (Gusfield et al. 2004a, 2004b), where the paths from the most recent common ancestor (MRCA) of the parents of a hybrid node down to the hybrid node form disjoint cycles; and (binary) "trees," which only contain tree nodes. These network classes are nested in this order: tree sibling ⊃ tree child ⊃ galled trees ⊃ trees, meaning a tree is also a galled tree, a tree child network, and a tree sibling network, and so on. Regular networks, however, are not related to tree sibling, tree child, or galled tree networks (Cardona G, Llabrés M, Rosselló F, Valiente G, Recent advances in metrics for phylogenetic networks, unpublished date).

The point we want to make is that the networks resulting from reticulating evolutionary processes do not necessarily correspond with any of the idealized classes of networks described above and for which perfect metrics exist. Were this true, it would imply that many phylogenetic networks could not be properly compared with the existing metrics. Obviously, it is very important to characterize the size of this perceived gap between the algorithms and the biology, and this is precisely our goal. In order to provide a formal statistical description of this disagreement, independent of particular organisms or genomic regions, we will use coalescent theory (Kingman 1982) to generate the type of reticulate networks that result from the evolutionary process. The coalescent describes the probabilities of the different genealogies for a sample of genes generally but not necessarily from the same population, and it was

Key words: phylogenetic network, reticulate evolution, coalescent.

E-mail: valiente@lsi.upc.edu.

*Mol. Biol. Evol.* 25(12):2517–2520. 2008
doi:10.1093/molbev/msn219
Advance Access publication October 15, 2008

further extended by Hudson (1983) to include recombination events. Specifically, we will quantify the different classes of networks (regular, tree sibling, tree child, and galled trees) produced by the coalescent as a function of the population recombination rate.

## Methods

### Simulation of Coalescent Networks

The standard coalescent describes the possible histories of a sample of genes back in time to their MRCA (Kingman 1982). In the absence of evolutionary forces like selection, migration, or recombination, the only types of events that can occur are coalescent events, in which two lineages (branches) fuse into one. Therefore, for a sample of $n$ genes, there will be $n - 1$ coalescent events until the MRCA is reached. If recombination is included, the recombination events will result actually in the opposite pattern going backward in time, as in this case one lineage (the recombinant) separates into two (the parents of the recombination event). Therefore, the genealogies produced by the coalescent with recombination (also known as ARG, for "ancestral recombination graph") will be explicit reticulate networks as internal nodes actually represent evolutionary events. In figure 1, we show a typical genealogy for a sample of 10 genes.

In our simulations, the genealogies were simulated with the program Recodon (Arenas and Posada 2007). Two sample sizes ($n$ = 10 and 50) and seven population recombination rates ($\rho$ = 0, 1, 2, 4, 8, 16, and 32) were explored. We used a continuous-time approximation where the times of the events are exponentially distributed. For every combination of parameters, we simulated 1,000 replicates, producing and evaluating thus a total of 14,000 genealogies. We varied the recombination rate that much because the recombination events actually determine the different network classes produced (regular, tree sibling, tree child, galled trees, or binary trees). The number of raw recombination events $R(n)$ in the network for a sample of $n$ genes has the expectation

$$E[R(n)] = \rho \sum_{j=1}^{n-1} \frac{1}{j},$$

and thus, 0, 2.72, 5.44, 10.87, 21.74, 43.49, and 86.97 recombination events are expected for $n$ = 10, and 0, 4.46, 8.92, 17.83, 35.67, 71.34, and 142.68 for $n$ = 50, for the recombination rates enumerated above, respectively.

### Assignment to Network Classes

The simulated genealogies were classified into five different network classes: regular, tree sibling, tree child, galled trees, and binary trees. First, hybrid and tree nodes in the graphs were identified as the recombinant and coalescent nodes in the genealogies, respectively, while superfluous nodes were eliminated (fig. 2). Superfluous nodes have just one parent and one child and result from concatenated recombination and coalescent events. Although they are real nodes (and therefore count for $R(n)$), they cannot be estimated from real data and, therefore, were removed from the networks before their classification.

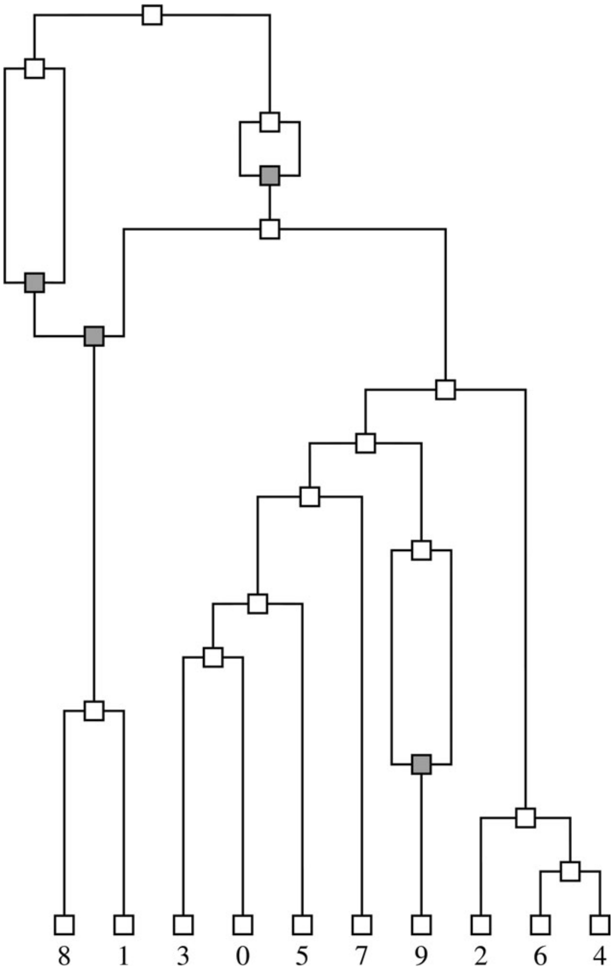

FIG. 1.—A single realization of the coalescent with recombination. The genealogy goes from the present (bottom) to the past (up). Coalescent and recombinant nodes are represented in white and gray, respectively.

## Results

In the absence of recombination, all the simulated genealogies were binary trees, as expected. With recombination, different classes of reticulate networks were produced whose complexity was a function of the population recombination rate (tables 1 and 2). For $n$ = 10, at small recombination rates ($\rho$ = 1) many networks were tree sibling, but half were already non-tree child. With twice as much recombination, half the networks were more complex than tree sibling networks, and only a few were tree child networks or galled trees. With moderate recombination rates ($\rho$ = 4), only 15% of the networks were tree sibling, whereas with larger recombination rates. almost no network, or even none at all, could be classified into any of the standard classes. Increasing the sample size to $n$ = 50 just increased a little more the complexity of the networks but preserved the same trend regarding the effect of recombination. On the other

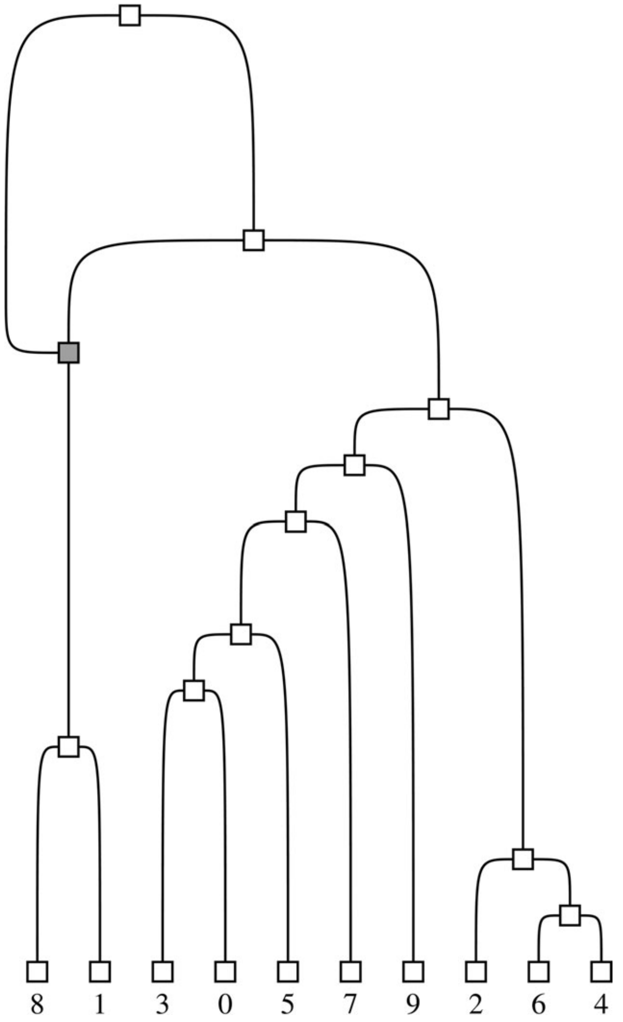

FIG. 2.—The corresponding graph for the ARG shown in figure 1. Tree and hybrid nodes are represented in white and gray, respectively. Note that superfluous nodes have been removed.

hand, even at small recombination rates ($\rho = 1$) rather few networks are regular because hybrid nodes in the genealogies simulated by the coalescent with recombination have two parents and a single child, and a network containing a hybrid node with a single child is not regular unless the child is also a hybrid node.

We will argue that the network topologies—we do not consider branch lengths—produced by the coalescent with recombination will be indistinguishable from those produced by other reticulating processes like gene conversion, hybridization, and lateral gene transfer. Note that the coalescent with recombination results, going backward in time, in nonreciprocal exchanges of genetic material as only one of the two recombinants inherits ancestral material and will be therefore represented in the network. Conveniently, these types of events are also typical of gene conversion, hybridization, and lateral gene transfer. In addition, the nodes in the ARG produced by the coalescent do not have to belong to the same species; they just represent gene copies. Therefore, we believe that these results are likely to be very similar for models of lateral gene transfer and/or hybridization—the assumptions underlying the standard classes of networks should eventually fail given sufficient reticulation.

We can conclude that the reticulate networks produced by the evolutionary process, at least as modeled by the coalescent with recombination, are much more convoluted than regular, tree sibling, tree child, or galled tree networks. These network classes—the only ones for which perfect metrics exist—are clearly insufficient to describe reticulating evolutionary processes. Indeed, new network metrics need to be developed if we really want to compare reticulate phylogenetic networks estimated from real data.

**Table 1**
**Number of Simulated Networks Falling in Each Class as a Function of the Recombination Rate $\rho$ = 0, 1, 2, 4, 8, 16, and 32 for Sample Size $n$ = 10**

| | Recombination Rate | | | | | | |
|---|---|---|---|---|---|---|---|
| Network Class | 0 | 1 | 2 | 4 | 8 | 16 | 32 |
| Regular | 1,000 | 200 | 58 | 5 | 0 | 0 | 0 |
| Tree sibling | 1,000 | 832 | 514 | 151 | 14 | 0 | 0 |
| Tree child | 1,000 | 560 | 205 | 39 | 1 | 0 | 0 |
| Galled trees | 1,000 | 440 | 137 | 21 | 1 | 0 | 0 |
| Trees | 1,000 | 139 | 27 | 1 | 0 | 0 | 0 |

**Table 2**
**Number of Simulated Networks Falling in Each Class as a Function of the Recombination Rate $\rho$ = 0, 1, 2, 4, 8, 16, and 32 for Sample Size $n$ = 50**

| | Recombination Rate | | | | | | |
|---|---|---|---|---|---|---|---|
| Network Class | 0 | 1 | 2 | 4 | 8 | 16 | 32 |
| Regular | 1,000 | 57 | 1 | 0 | 0 | 0 | 0 |
| Tree sibling | 1,000 | 784 | 469 | 101 | 2 | 0 | 0 |
| Tree child | 1,000 | 463 | 126 | 9 | 0 | 0 | 0 |
| Galled trees | 1,000 | 161 | 5 | 0 | 0 | 0 | 0 |
| Trees | 1,000 | 34 | 0 | 0 | 0 | 0 | 0 |

## Acknowledgments

We thank the editor and an anonymous reviewer for their comments. M.A. was supported by an FPI fellowship BES-2005-9151 (Spanish Ministry of Education and Science [MEC]). G.V. was partially supported by MEC and EU FEDER funds through grants MTM2006-07773 COMGRIO and MTM2006-15038-C02-01. D.P. was partially supported by MEC and EU FEDER funds through grant BIO2007-61411.

## Literature Cited

Arenas M, Posada D. 2007. Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography. BMC Bioinformatics. 8:458.

Baroni M, Semple C, Steel M. 2004. A framework for representing reticulate evolution. Ann Combin. 8:391–408.

Cardona G, Llabrés M, Rosselló F, Valiente G. 2008a. A distance metric for a class of tree-sibling phylogenetic networks. Bioinformatics. 24(13):1481–1488.

Cardona G, Llabrés M, Rosselló F, Valiente G. Forthcoming 2008b. Metrics for phylogenetic networks I: generalizations of the Robinson-Foulds metric. IEEE/ACM Trans Comput Biol Bioinform.

Cardona G, Rosselló F, Valiente G. 2008a. A perl package and an alignment tool for phylogenetic networks. BMC Bioinformatics. 9:175.

Cardona G, Rosselló F, Valiente G. 2008b. Comparison of tree-child phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform. doi: 10.1109/TCBB.2007.70270.

Cardona G, Rosselló F, Valiente G. 2008c. Tripartitions do not always discriminate phylogenetic networks. Math Biosci. 211(2): 356–370.

Cassens I, Mardulyn P, Milinkovitch MC. 2005. Evaluating intraspecific "network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? Syst Biol. 54(3):363–372.

Cassens I, van Waerebeek K, Best PB, Reyes J, Milinkovitch MC. 2003. The phylogeography of dusky dolphins (*Lagenorhynchus obscurus*): a critical examination of network methods and rooting procedures. Mol Ecol. 12(7):1781–1792.

Gusfield D, Eddhu S, Langley C. 2004a. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. J Bioinform Comput Biol. 2(1):173–213.

Gusfield D, Eddhu S, Langley C. 2004b. The fine structure of galls in phylogenetic networks. INFORMS J Comput. 16(4): 459–469.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 23(2):183–201.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2): 254–267.

Jin G, Nakhleh L, Snir S, Tuller T. 2007. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Mol Biol Evol. 24(1):324–337.

Kingman JFC. 1982. The coalescent. Stoch Process Appl. 13(3):235–248.

Moret BME, Nakhleh L, Warnow T, Linder CR, Tholse A, Padolina A, Sun J, Timme R. 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM Trans Comput Biol Bioinform. 1(1):13–23.

Posada D, Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol. 16(1): 37–45.

Woolley SM, Posada D, Crandall KA. 2008. A comparison of phylogenetic network methods using computer simulation. PLoS ONE. 3(4):e1913.

Asger Hobolth, Associate Editor

Accepted September 23, 2008