Phylogenetic Networks

Gary A. Churchill
Biometrics Unit
Cornell University
Ithaca, NY 14853

BU-1297-M

June 1995

Phylogenetic Networks

Gary A. Churchill

1 Introduction

It is a generally accepted principle in molecular evolution that relationships among gene sequences sampled from a group of organisms can be represented as a bifurcating tree. Indeed phylogenetic trees almost always provide an adequate representation for the history of sequences sampled from moderately divergent species. However there are some situations for which a more general representation is required. In this note, network phylogenies are proposed as a graphical summary of relationships among a set of sequences.

Inference procedures for network phylogenies have been described by von Haeseler and Churchill (1993) for the case of binary character states and a symmetric model of character change. Their model, summarized below, is based on a multinomial likelihood for the observed character states configurations. A distance based method for inferring networks has been described by Bandelt and Dress (1990) and applied to viral sequences by Dopazo et al. (1990).

2 The Network Model

Suppose we have a set of K binary sequences each of length N with homologous sites aligned in columns. Thus the data are a matrix with elements $s_{ij} \in \{0,1\}$ where i is the species index and j is the site index. Each site j forms a K–dimensional vector of zeroes and ones, a binary character–state configuration. We assume that the evolutionary process acts independently and identically at each site and that changes between the states zero and one are symmetric. Thus two configuration are equivalent if one can be obtained from the other by reversing the labeling of the zero and one states.

A split is a bipartition of the K species into two disjoint sets and can be represented by the set S, such that $1 \in S$, and the set $S^c = \{1,\ldots,K\} \setminus S$. Each of the 2^{K-1} non–equivalent configurations corresponds to a split on the species. A network model $M = (\mathcal{S}, \theta)$ is defined by a set of splits and a parameter vector $\theta = \{\theta_S | S \in \mathcal{S}\}$, where θ_S is the probability that, at a given site, an odd number of substitutions separate the sequences in S from those in S^c. The parameter vector θ is analogous to Hendy and Penny's (1989) $\mathbf{p}$ vector, the probability of character changes along an edge of a tree.

A network model is a tree if and only if all splits in $\mathcal{S}$ are pairwise compatible. Splits S_i and S_j are pairwise compatible if there exists a $B_i \in \{S_i, S_i^c\}$ and a $B_j \in \{S_j, S_j^c\}$ with $B_i \cap B_j = \emptyset$ (Buneman (1971)). For example, consider the model M_1, with $\mathcal{S} = \{S_1, S_2, S_3, S_4\}$, where $S_1 = \{1\}, S_2 = \{1,3,4\}, S_3 = \{1,2,4\}, S_4 = \{1,2,3\}$,. This is the familiar four species star phylogeny. A new model M_2 is created by adding the split $S_5 = \{1,2\}$. This is the four species tree that pairs species 1 and 2 versus species 3 and 4. If we now obtain a third model M_3 by adding the split $S_6 = \{1,3\}$ the result is a network. The splits S_5 and S_6 are not compatible.

3 Applications

It is important to make a distinction between networks as a represenation of statistical uncertainty in the inference of a true tree-like phylogeny and networks as representation of an underlying biological phenomenon.

Statistical Networks. It is often the case that there is insufficient information in a sequence set to determine a unique and fully bifurcating tree with certainty. The collection of trees consistent with a dataset, in the sense that each cannot be rejected by a hypothesis test, consititutes a confidence interval (Cavender 1978, Felsenstein 1988). Unfortunately such collections may be large and hard to visualize. A set of trees can be combined by taking the union of all splits in the trees and forming the corresponding network. Conversely, a network can be decomposed into a set of trees each composed of a maximal set of compatable splits. A minimal network that contains the set of trees compatable with the data is a confidence set with a compact representation as a set of splits. The network confidence set is more precise than than a multifurcating tree.

Biological Networks. Not all relationships among sets of sequences are treelike. First, there are increasing numbers of examples of horizontal transfers of genetic information across species boundaries (reviewed by Syvanen, 1994) including evidence for ancient fusions of whole genomes (Golding and Gupta, 1995). Second, the events that take place at the bifurcation in a phylogenetic tree can be very complex. Population subdivision with some gene flow and/or hybridization can result in gene trees that do not agree with species trees. Third, when we are looking at sequence derived from within a species, different segments of a sequence are likely to have unique histories due to recombination. Ideally we would like to identify the recombination breakpoints and assign a

tree–like history to each segment. However, the amount of sequence variation may limit our ability to make such inferences. This list of situations that would give rise to network relationships is not exhaustive.

4 Alternative Models

The network model allows multiple events on non–compatable splits to act on the same site in a sequence. This approach contradicts the notion that each site should have a tree–like history. Alternative models could be developed if we insist on tree–like evolution at each site but wish to allow the tree to vary from site to site.

Mixture Model. In a mixture model, each site in the aligned sequence set is assumed to have an unknown phylogenetic tree that is drawn from a finite set $\mathcal{T} = \{\tau_1, \ldots, \tau_k\}$ of possible trees. The tree $\tau_{k(i)}$ corresponding to site i is drawn independently with probability $\alpha_{k(i)}$ from the set $\mathcal{T}$. If the set of possible trees is known, the resulting likelihood is a multinomial mixture and can be maximized using standard methods for mixture estimation. Estimation of the set $\mathcal{T}$ represents a much more challenging problem. A reasonable starting point might be obtained by fitting a network model and taking $\mathcal{T}$ to be the set of subtrees of the network.

Block Models. When a small number of recombination or gene conversion events have occurred in the history of the sequence set, blocks of adjacent sites will share identical tree–like histories. Sawyer's (1989) test for gene conversion is based on detection of runs of compatable configurations. The problem of reconstructing the trees and the recombination break points has been studied by Hein (1990) using parsimony methods. The model described by Hein is implicitly a hidden Markov model. Related applications of hidden Markov models to sequence data have been described by Churchill (1989) and by Felsenstein and Churchill (1995). The hidden Markov model assumes that topology changes along the aligned sequence set form a Markov chain. Unfortunately, segments generated by standard models of recombination are not Markovian. Robust properties of the hidden Markov model suggest that it may provide a tractable approximation to more realistic block models.

Rate Heterogeneity. Variation in substitution rates between sites can induce an apparent network structure in sequence data. When rate heterogeneity is present, estimated rates assuming a homogeneous model will be an average

of the actual rates. Sites with a high rate of change are likely to show an apparent excess of parallel changes in independent lineages. The introduction of invariable sites (Shoemaker and Fitch, 1989) into the network model will reduce the number of statistically significant non–compatable splits. More general models of rate heterogeneity should be considered in conjuction with a network phylogeny. In principle, it should be possible to distinguish rate heterogenity from network structure as heterogeneity should inflate the significance of all splits equally.

5 Summary

Network models extend the classical model of phylogeny (e.g. Felsenstein 1981) by allowing for a more general dependence structure among gene sequences. Networks can be interpreted statistically as mixtures of tree–like models and can be useful in modeling a variety of evolutionary scenarios. Further work is needed to develop mixture models, block models, models of rate heterogeneity and associated methods for statistical analysis of network relationships among sequences.

References

[1] Bandelt HJ, Dress AWM (1990) A canonical decomposition theory for metrics on a finite set. Preprint 90–032, SFB 343, Universität Bielefeld.

[2] Buneman P (1971) The recovery of trees from measures of dissimilarity. In: Hodson FR, Kendall DG, Tantu P (eds) Mathematics in the archaeological and historical science. Proc. of the Anglo–Romanian–Conference 1970. Univ. Press Edinburgh. pp 387–395.

[3] Cavender J (1978) Taxonomy with confidence. Math. Biosc. 40:271–280.

[4] Churchill GA (1989) Stochastics Models for heterogeneous DNA sequences Bull. Math. Biol.

[5] Dopazo J, Dress A, von Haeseler A (1990) Split decomposition: A new technique to analyze viral evolution. Preprint 90–037, Sonderforschungsbereich 343 Diskrete Strukturen in der Mathematik. Universität Bielefeld.

[6] Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

[7] Felsenstein J (1988) Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22:521–565.

[8] Felsenstein J, Churchill GA (1995) Mol. Biol. Evol. (submitted)

[9] Golding B, Gupta (1995) Mol. Biol. Evol.

[10] von Haeseler A, Churchill GA (1993) Network models for sequence evolution. J. Mol. Evol. 37:77–85.

[11] Hein J (1990) reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. 98:185–200.

[12] Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

[13] Sawyer S (1989) Statistical test for detecting gene conversion. Mol. Biol. Evol. 6:526–538.

[14] Shoemaker JS, Fitch WM (1989) Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270–289.

[15] Syvanen M (1994) Horizontal gene transfer: Evidence and possible consequences. Ann. Rev. Genet. 28:237–261.