A Survey of Combinatorial Methods for Phylogenetic Networks

Daniel Huson, Celine Scornavacca

Genome Biology and Evolution

A Survey of Combinatorial Methods for Phylogenetic Networks

Daniel H. Huson*,† and Celine Scornavacca*,†

Department of Computer Science, Center for Bioinformatics (ZBIT), Tübingen University, Tübingen, Germany

*Corresponding author: E-mail: scornava@informatik.uni-tuebingen.de.

†These authors contributed equally to this work.

Accepted: 11 November 2010

Abstract

The evolutionary history of a set of species is usually described by a rooted phylogenetic tree. Although it is generally undisputed that bifurcating speciation events and descent with modifications are major forces of evolution, there is a growing belief that reticulate events also have a role to play. Phylogenetic networks provide an alternative to phylogenetic trees and may be more suitable for data sets where evolution involves significant amounts of reticulate events, such as hybridization, horizontal gene transfer, or recombination. In this article, we give an introduction to the topic of phylogenetic networks, very briefly describing the fundamental concepts and summarizing some of the most important combinatorial methods that are available for their computation.

Key words: networks, reticulate events, phylogeny, molecular, combinatorics.

Introduction

Phylogenetic analysis aims at uncovering the evolutionary relationships between different species or taxa in order to obtain an understanding of the evolution of life on Earth. "Phylogenetic trees" are widely used to address this task and are usually computed from molecular sequences. By definition, phylogenetic trees are well suited to represent evolutionary histories in which the main events are speciations (at the internal nodes of the tree) and descent with modification (along the edges of the tree).

However, these trees are less suited to model mechanisms of "reticulate evolution" (Sneath 1975), such as horizontal gene transfer, hybridization, recombination, or reassortment. Moreover, mechanisms such as incomplete lineage sorting or complicated patterns of gene duplication and loss can lead to incompatibilities that cannot be represented on a tree. Although the analysis of individual genes or short stretches of genomic sequences often supports a single phylogenetic tree, different genes, or sequence segments usually support different trees.

"Phylogenetic networks" provide an alternative to phylogenetic trees when analyzing data sets whose evolution involves significant amounts of reticulate events (Sneath 1975; Syvanen 1985; Delwiche and Palmer 1996; Griffiths and Marjoram 1996; Rieseberg 1997; Doolittle 1999). Moreover, even for a set of taxa that have evolved according to a tree-based model of evolution, phylogenetic networks can be usefully employed to explicitly represent conflicts in a data set that may be caused by mechanisms such as incomplete lineage sorting or by the inadequacies of an assumed evolutionary model (Huson and Bryant 2006).

Although rooted phylogenetic networks can, in theory, be used to explicitly describe evolution in the presence of reticulate events, their calculation is difficult and computational methods for doing so have not yet matured into practical and widely used tools (Hein 1993; Gusfield et al. 2003; Huson et al. 2005; Song et al. 2005; Bordewich et al. 2007; Tofigh et al. 2010). In contrast, a number of established tools for computing unrooted phylogenetic networks can be used to visualize incompatible evolutionary scenarios in phylogeny and phylogeography (Bandelt and Dress 1992; Bandelt et al. 1995, 1999; Huson 1998; Clement et al. 2000; Bryant and Moulton 2004; Huson and Bryant 2006).

In this paper, we give an introduction to the topic of phylogenetic networks, very briefly describing the fundamental concepts and summarizing some of the most important methods that are available for the computation of phylogenetic networks. In practice, most currently available algorithms for computing phylogenetic networks are based on combinatorics, so we focus on these approaches. Some approaches developed within a maximum parsimony or

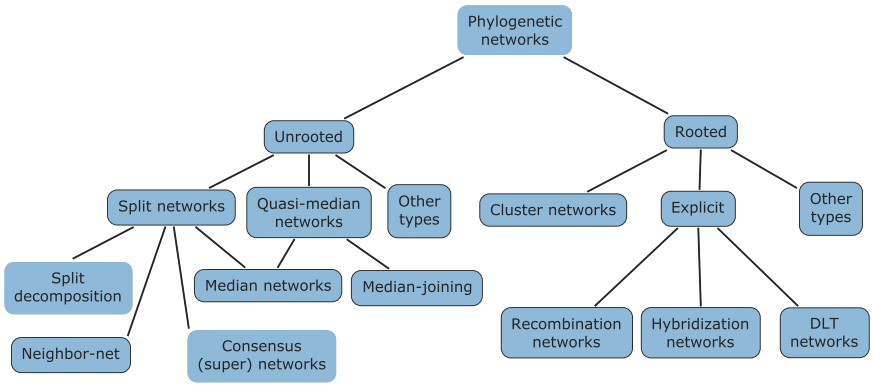

Fig. 1.—Overview of the main concepts mentioned in this paper. First, we distinguish between unrooted (on the left) and rooted networks (on the right). Although all phylogenetic networks mentioned on the left generalize unrooted phylogenetic trees, all those mentioned on the right generalize rooted trees. Second, we distinguish between explicit networks (shown below the node labeled Explicit on the right) and abstract ones (all others).

maximum likelihood framework can be found, for example, in Hein (1993); Jin et al. (2006a, 2006b, 2007); Dessimoz et al. (2008). Figure 1 shows the relationships between some of the concepts mentioned in this paper. For ease of exposition, some of the more technical terms in this survey are defined in table 1.

The purpose of this paper is to give a short survey of the combinatorial methods used to infer phylogenetic networks. More details on the concepts and algorithms introduced in this paper as well as biological examples of their applications can be found in (Huson et al. 2011).

What is a Phylogenetic Network?

In the literature, the term phylogenetic network is defined and used in a number of different ways, usually focusing on the specific type of network that an author happens to be interested in (Bandelt 1994; Gusfield et al. 2003; Linder and Rieseberg 2004). We propose the following general definition:

Definition 1 (Phylogenetic network) A phylogenetic network is any graph used to represent evolutionary relationships (either "abstractly" or "explicitly"; see below) between a set of taxa that label some of its nodes (usually the leaves).

Phylogenetic networks can be computed from a wide range of data, including multiple sequence alignments, distance matrices, set of trees, clusters, splits, rooted triplets, or unrooted quartets. As with phylogenetic trees, a first major distinction is between "unrooted" and "rooted" phylogenetic networks:

Definition 2 (Unrooted phylogenetic network) Let X be a set of *taxa*. An "unrooted phylogenetic network" N on X is any undirected graph whose leaves are bijectively labeled by the taxa in X.

A number of different types of unrooted phylogenetic networks are in use. In this paper, we mainly focus on the important class of "split networks" (Bandelt and Dress 1992). A second important class of unrooted phylogenetic networks are "quasi-median networks," which can be viewed as a generalization of split networks.

A "rooted Direct Acyclic Graph (DAG)" is a directed graph that is free of directed cycles and that contains precisely one node without ancestors, called the "root." Rooted phylogenetic networks generalize rooted phylogenetic trees:

Definition 3 (Rooted phylogenetic network). Let X be a set of taxa. A "rooted phylogenetic network" N on X is a rooted DAG where the set of leaves is bijectively labeled by the taxa in X.

For an example of unrooted and rooted phylogenetic networks, see figure 2.

The envisioned role of rooted phylogenetic networks in biology is to describe the evolution of life in a way that explicitly includes reticulate events. Ultimately, the main goal is to work out the details of a rooted phylogenetic network of life, such as popularized by Doolittle (1999).

Phylogenetic networks can be used in two different ways. The first use is as a tool for visualizing incompatible data sets in a helpful manner, in which case we speak of an "abstract" phylogenetic network. The second type of usage is as a representation of a putative evolutionary history involving reticulate events, in which case, the network is called "explicit."

By definition, most (if not all) types of unrooted phylogenetic networks are abstract networks, as evolution is inherently rooted (and thus any unrooted phylogenetic tree is also abstract, in this sense). However, rooted phylogenetic

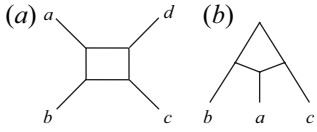

Fig. 2.—(*a*) An unrooted phylogenetic network on X = {a, b, c} *d* and (*b*) a rooted phylogenetic network on X = {a, b, c} in which the top node is the root.

Table 1
Terms Used in the Text without Definition

Biconnected component	**A graph that consists of only one node or of two nodes joined by a single edge or that has more than two nodes and any two nodes *v*, *w* are connected by at least two different paths that are node disjoint (except at *v* and *w*).**
Circular split set	**A set of splits that can be represented by an outer-labeled planar split network.**
Compatible split set	**A set of splits that can be represented by an unrooted phylogenetic tree.**
Condensed version of *M*	**A multiple sequence alignment $\bar{M}$ obtained from *M* by deleting sequences and columns such that no two sequences are identical, no two columns induce the same partitioning, and no constant columns are present.**
Cluster	**A proper subset of a set of taxa.**
Outer-labeled planar graph	**A graph that can be drawn in the plane such that no two edges intersect and all labeled nodes lie on the outside of the graph.**
Quasi-median of three sequences *a*, *b*, *c* of length *L*	**The set qm (*a*, *b*, *c*) of all sequences $d = d_1 \ldots d_L$ that have the property that the state d_i occurs in the set $\{a_i, b_i, c_i\}$ at least as many times as any other state, for each position $i = 1, \ldots, L$.**
Taxon	**A taxonomic unit that represents a group of organisms.**
Split	**A bipartition of a set of taxa, for example, induced by an edge of an unrooted phylogenetic tree.**

networks can be either abstract or explicit, depending on how they are constructed and interpreted.

The necessity of distinguishing between abstract and explicit networks was pointed out in Morrison (2005). They are called implicit and explicit in Huson (2007). In Morrison (2010), abstract and explicit networks are named "data-display" networks and "evolutionary" networks, respectively.

In the literature, perhaps as many as 20 different names have been defined for different types of phylogenetic networks. A closer look reveals that some networks are named by the algorithms that compute them or by mathematical properties that define them, such as "neighbor-nets" or "median networks." Others are named by topological constraints that are imposed on them for computational reasons, such as "galled trees," "galled networks," or "level-*k* networks." Yet others are named by the types of evolutionary events which they model, such as "hybridization networks," "recombination networks," or "duplication-loss-transfer (DLT) networks."

Unrooted Phylogenetic Networks

A number of special types of unrooted phylogenetic networks are used in practice, the most important of which we consider in detail in this paper.

Split Networks

The foundation for split networks was laid in Bandelt and Dress (1992). Let X be a set of taxa and assume that we are given a set of "splits" $\mathcal{S}$ on X, usually with a "weighting" that assigns a nonnegative weight to each split, which may represent character changes or distances or may also have a more abstract interpretation. If the set of splits $\mathcal{S}$ is "compatible," then it can be represented by an unrooted phylogenetic tree, and each edge in the tree corresponds to exactly one of the splits (Buneman 1971). More generally, S can always be represented by a "split network," which is an unrooted phylogenetic network with the property that every split S in $\mathcal{S}$ is represented by an array of parallel edges in N.

An example is shown in figure 3, where the three central edges highlighted in bold represent the split that separates the outgroups from the Branchiopoda. Indeed, the removal of these edges produces precisely two subtrees, one which has leaves that are labeled by Branchiopoda species and the other with leaves that are labeled by outgroup species.

Two methods for constructing split networks from weighted splits are the "convex hull algorithm" and the "circular network algorithm." The convex hull algorithm can be applied to any set of splits $\mathcal{S}$ and computes a split network representing $\mathcal{S}$ that contains an exponential number of nodes and edges in the worst case (Bandelt et al. 1995). It is also used to compute median networks, as described below. The circular network algorithm can be applied to any set of "circular splits" and produces an "outer-labeled planar" network with only a quadratic number of nodes and edges (Dress and Huson 2004).

In many cases, direct application of the convex hull algorithm leads to an overcomplicated network. In practice, a useful heuristic is first to choose an order of the taxa such that a large subset of the given set of splits is circular. This subset of splits is then processed using the circular network algorithm to obtain an outer-labeled planar network. The remaining splits are then processed using the convex hull algorithm, which will add some nonplanar parts to the network.

A split network N can be obtained from a number of different types of data. To be more precise, the algorithms mentioned below do not compute a split network directly; rather, they all compute a set of weighted splits $\mathcal{S}$. A split network N is then computed from $\mathcal{S}$ as described above. All splits-based algorithms discussed in this article are implemented in the program SplitsTree4 (Huson and Bryant 2006).

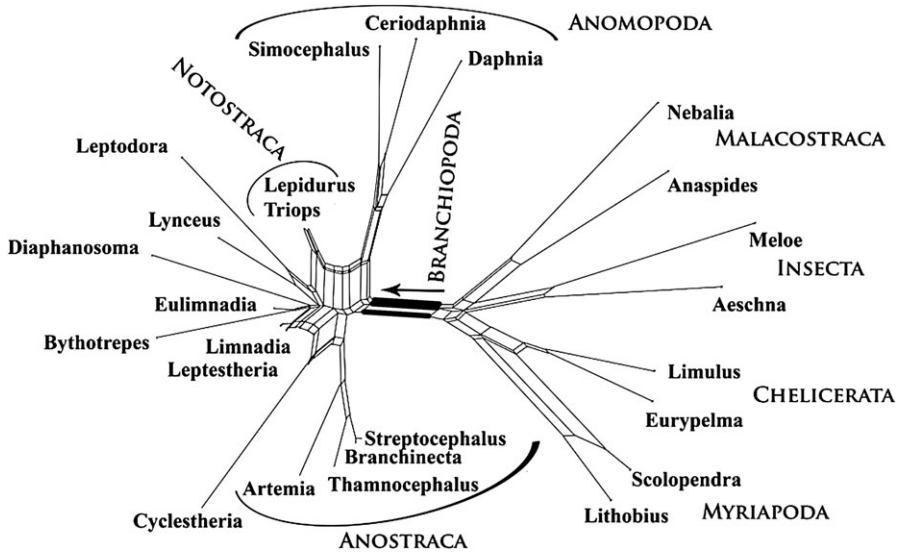

Fig. 3.—A split network on 25 species of Branchiopoda and outgroups, computed from 18S rDNA sequences using Neighbor-Net, as reported in Wagele and Mayer (2007). The authors compare this network with a maximum parsimony tree for the same data set and discuss how the network exhibits conflicting signals that are not represented in the tree. Reprinted from BMC Evolutionary Biology 7:147 (2007) under the Creative Commons Attribution License.

Split Networks from Distances

A number of methods exist for computing a set of weighted splits for a given distance matrix D on X. The two most important are split decomposition (Bandelt and Dress 1992) and "Neighbor-Net" (Bryant and Moulton 2004).

"Split decomposition" takes a distance matrix D on X as input and produces a set of weighted splits $\mathcal{S}$ on X that is "weakly compatible," a property that ensures that the corresponding split network will not be too complicated. Indeed, in practice, the resulting split networks are often quite close to being outer-labeled planar, as they usually have only a few edges crossing over each other and do not contain any "high-dimensional cubes," which may occur for completely unrestricted sets of splits. In practice, split decomposition is a very conservative method, in the sense that a split will only be present in the output if there is global support for it in the given data set. For large or diverse data sets, the method tends to exhibit very low resolution and thus its use is limited to small data sets of less than 100 taxa, say.

Neighbor-Net takes a distance matrix D on X as input and produces a set of weighted splits $\mathcal{S}$ on X that is circular and can be represented by a outer-planar split network using the circular network algorithm. Neighbor-Net is more popular than split decomposition because it is less conservative and so does not lose resolution on larger data sets. Moreover, the fact that the output of the method can always be represented by a outer-planar split network and is thus easy to visualize adds to its attraction; see figure 3.

Both network methods have the attractive property that they produce the set of splits corresponding to the correct tree when given a tree-like matrix.

Split Networks from Trees

Let $\mathcal{T} = (T_1, \ldots, T_k)$ be a collection of unrooted phylogenetic trees on X. These might be different gene trees, trees for the same gene computed using different methods, or a set of trees obtained in a Bayesian analysis, for example. Split networks can be used to visualize conflicting signals present in $\mathcal{T}$.

The set of majority-consensus splits is defined as the set of all splits that are present in more than 50% of the input trees. By lowering the threshold to a proportion p of 50% or less, one obtains a set of splits $\mathcal{S}_p(\mathcal{T})$ that will not necessarily be compatible. The split network N associated with $\mathcal{S}_p(\mathcal{T})$ is called a "consensus split network" and can be used to visualize conflicting signals in a set of trees (Holland and Moulton 2003).

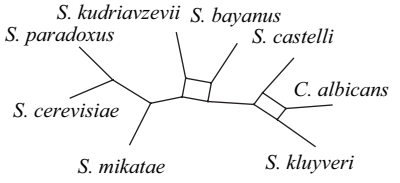

Fig. 4.—For a set T of 106 phylogenetic trees on eight yeast species reported in Rokas et al. (2003), we show the consensus split network representing all splits that occur in more than 30% of the trees.

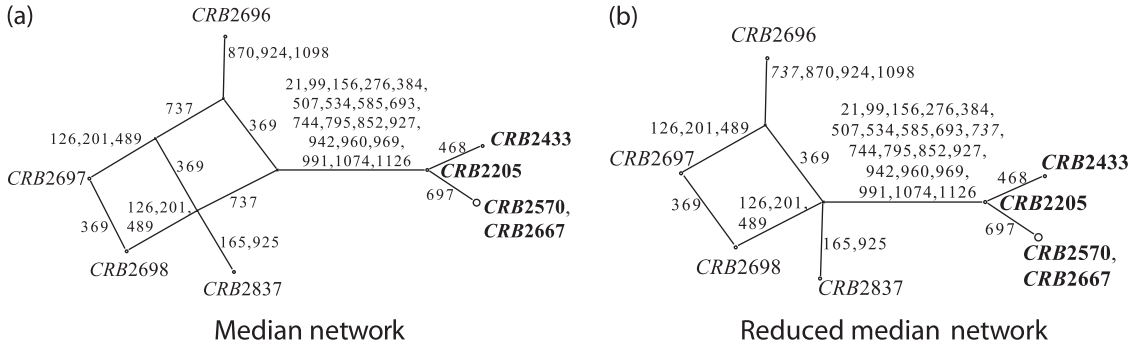

Fig. 5.—(*a*) The median network for eight specimens of *Callicebus lugens*, based on cytochrome *b* sequences (data from Casado et al. [2007]). Specimens from the right bank of Rio Negro are shown in plain font and those from the left bank are shown in bold font. (*b*) The reduced median network obtained by postulating a parallel mutation at position 737.

One of the first published applications of this method was to a collection of 106 different unrooted phylogenetic trees involving eight different yeast species (Holland et al. [2004], gene trees from Rokas et al. [2003]).

Figure 4 shows clearly that the gene trees disagree somewhat as to where the outgroup taxon *Candida albicans* attaches to the phylogeny. Moreover, they also disagree on whether *Saccharomyces kudriavzevii* and *Saccharomyces bayanus* are sister taxa.

In practice, in a collection of gene trees, the set of taxa that occurs in each tree will often differ between trees, simply because some gene sequences may not be available for all taxa. To address this, methods have been developed to compute a "super split network" for a given set of unrooted phylogenetic trees T on overlapping but nonidentical taxon sets using the "Z-closure" algorithm (Huson et al. 2004; Whitfield et al. 2008).

Split Networks from Sequences

Assume that we are given a multiple sequence alignment M on X.

A first approach to obtaining a split network for M is to compute a set of splits that represents M using the "parsimony-splits" method (Bandelt and Dress 1993). This method takes a multiple alignment M on X as input and produces a set of weakly compatible splits $\mathcal{S}$ on X using a simple modification of the split decomposition algorithm. The parsimony-splits method has not been used much in the literature, probably because the resulting set of splits is usually very similar to the one obtained by the more widely known split decomposition.

Another way of computing a split network from M is first to restrict M to obtain a matrix $\hat{M}$ containing only the columns in M that contain exactly two different character states and then focus on the "condensed version" of $\hat{M}$, say $\bar{M}$. Then, any column of $\bar{M}$ defines a different split of the taxon set, and the set of splits $\mathcal{S}(\bar{M})$ obtainable in this way can be represented by a split network N. If we label each edge of the network N by the columns in the alignment $\bar{M}$ that correspond to the split represented by the edge, then the resulting split network is called a median network (Bandelt et al. 1995). This construction is suitable for data sets that have very few differences in them. Hence, median

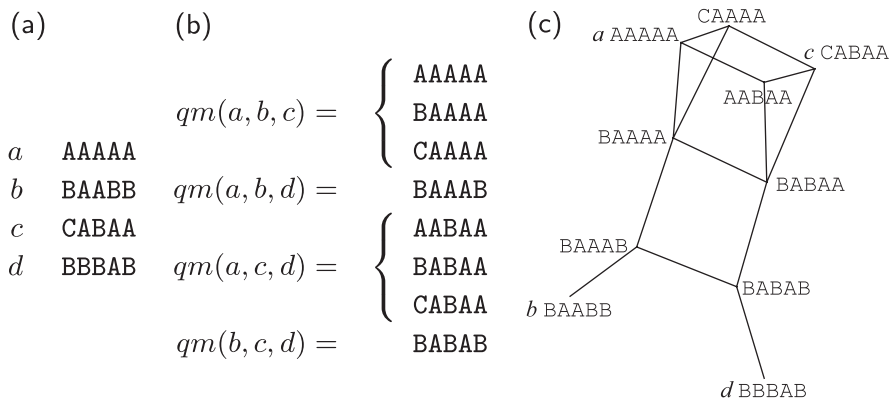

Fig. 6.—(*a*) A multiple condensed sequence alignment M. The quasi-median closure of M consists of the sequences depicted in (*a*) and (*b*). (*c*) The corresponding quasi-median network N.

networks are mainly used in phylogeography and population studies. Because parallel mutations can lead to complicated structures in such a network, the concept of a "reduced" median network was also introduced (Bandelt et al. 1995), in which one attempts to simplify the network by postulating appropriate parallel mutation events.

An example of a median network is shown in figure 5. In Casado et al. (2007), the distribution of *Callicebus lugens* (Platyrrhini, Primates) at the Rio Negr, in Brazil, is reported. The study focuses on eight specimens, one group of four taken from the left bank of the river and another group of four taken from the right bank. It is based on a multiple alignment M of cytochrome b DNA sequences of length 1,140. A median-network analysis shows a clear separation of the two groups. Note that only 35 columns are retained in the condensed version of M (the ones labeling the edges).

Split Networks from Quartets

Mathematicians are interested in developing methods that infer a phylogenetic tree or network from basic building blocks. In the computation of an unrooted tree or split network, these are phylogenetic trees on sets of four taxa, sometimes called "quartet trees." One such method is the "quartet-net" method, or "QNet," for short (Grünewald et al. 2007). This algorithm takes a set Q of weighted quartet topologies on X as input and, using a modification of Neighbor-Net, produces a set of weighted splits $\mathcal{S}$ on X that is circular, and thus can be represented by an outer-planar split network. Because compatible splits are always circular, it follows that the QNet method (combined with the circular network algorithm) always computes the correct phylogenetic tree when given an input set that corresponds to a tree.

Quasi-Median Networks

As we mentioned in the previous section, a median network can be used to visualize a set of binary characters on a set of taxa X. The concept of a quasi-median network is a generalization of the concept of a median network that was introduced to represent multistate characters. Note that, unlike median networks, quasi-median networks are not split networks. A quasi-median network is defined as a phylogenetic network, the node set of which is given by the quasi-median closure of the condensed version of M and in which any two nodes are joined by an edge if and only if the sequences associated with the nodes differ in exactly one position. The quasi-median closure is defined as the set of all sequences that can be obtained by repeatedly taking the "quasi-median of any three sequences" in the set and then adding the result to the set (see fig. 6).

In general, the quasi-median closure consists of a huge set of sequences and hence the quasi-median network for a multiple sequence alignment M of DNA sequences on X is usually too large and too complicated to be of practical interest. At the other extreme, a "minimum spanning network" (Excoffier and Smouse 1994; Bandelt et al. 1999) can be used to represent the differences between the sequences in M. This type of network is also often of limited interest because it contains one node per taxon and no additional nodes.

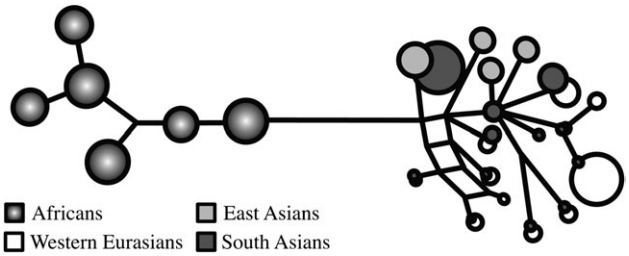

Fig. 7.—A median-joining network on human populations computed from mtDNA (adapted from Kivisild et al. [1999]; Disotell [2003]). Each disk in the tree represents a cluster of human mitochondrial types, and its diameter is proportional to the number of sequences represented. For African sequences, edges between individual types are collapsed and not shown. The cluster containing all non-African sequences is shown here in a noncollapsed view.

The "median-joining" algorithm (Bandelt et al. 1999) constructs an informative subnetwork of the full quasi-median network, repeatedly using the concept of a (relaxed) minimum spanning network and repeatedly employing the quasi-median calculation. Although the former construction, on its own, will produce too few nodes to be useful, the latter construction alone will produce too many nodes. By using both together, the median-joining method attempts to provide a useful network of intermediate size. The median-joining method is best suited for closely related sequences that have evolved without recombinations and is widely used in phylogeography and population studies, usually based on mtDNA or the Y chromosome. An application is shown in figure 7, where the cluster containing all non-African sequences attaches to only one of the clusters of African lineages. This network is thus consistent with the out-of-Africa model of human origins, suggesting that all non-African populations are derived from one African lineage.

Implementations of the median network and median-joining algorithms are provided by the programs Network (http://www.fluxus-engineering.com) and SplitsTree4.

Other Types

A number of other types of unrooted phylogenetic networks are in use. We briefly describe two of them.

Haplotype Networks

A haplotype network is an unrooted phylogenetic network in which the nodes represent different haplotypes within a group of (usually very closely related) taxa and the edges

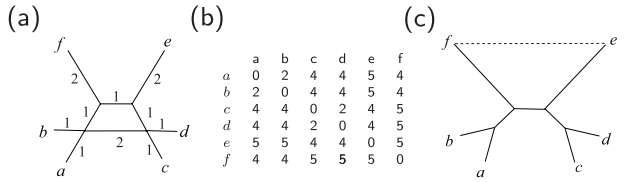

Fig. 8.—(*a*) A reticulogram R on $X = \{a, \ldots, e\}$ with edges labeled by their lengths. (*b*) The distance matrix D_R on X that is defined by R. (*c*) The reticulogram $R_{\mathrm{TRex}}(D_R)$ obtained by applying the T-Rex algorithm to D_R. Solid lines represent the initial unrooted phylogenetic tree, and a dashed line indicates the added shortcut edge.

join those sequences or haplotypes that are very similar. The edges are usually labeled by the positions at which the joined haplotypes differ.

Both the median network computation and the median-joining algorithm can be used to compute a haplotype network. Another popular approach is the "TCS approach" (Templeton et al. 1992). It is based on the concept of statistical parsimony and aims at producing a haplotype network in which two haplotypes are joined by an edge if and only if a quantity called the "probability of parsimony" (defined in Templeton et al. [1992], eqs. 6–8) exceeds 95% for the edge. The TCS method is similar to the (quasi-)median network method in that it attempts to place sequences onto the nodes of a network, infer additional nodes and label edges by the number of differences between different sequences. An implementation is available from: http://darwin.uvigo.es/software/tcs.html.

Reticulograms

A reticulogram is an unrooted phylogenetic tree to which a set of auxiliary edges has been added. A reticulogram is obtained from a distance matrix D on X using the T-Rex software, which first computes a phylogenetic tree on X (using a method such as neighbor-joining) and then repeatedly adds shortcut edges to the graph until the distances between the taxa in the graph show a good fit to the distances in the original input matrix D (Makarenkov 2001). An implementation is available from: http://www.trex.uqam.ca.

Unfortunately, it is easy to construct a reticulogram R on X such that the T-Rex algorithm will fail to reconstruct R from the distance matrix D_R (see fig. 8).

Rooted Phylogenetic Networks

Let X be a set of taxa and N a rooted phylogenetic network on X. Any node of indegree ≥ 2 is called a "reticulate" node and all others are called "tree" nodes. Any edge leading to a reticulate node is called a reticulate edge and all others are called tree edges. Definition 3 is very general and additional requirements can be made. For example, the network can be described as "bicombining," that is, that all reticulate nodes have indegree 2.

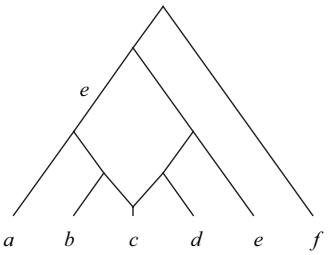

Fig. 9.—A rooted phylogenetic network N in which the edge e represents the cluster $\{a, b, c\}$ in the hardwired sense and the two clusters $\{a, b\}$ and $\{a, b, c\}$ in the softwired sense. Note that N does not represent the cluster $\{a, b\}$ in the hardwired sense.

How do we interpret such a rooted phylogenetic network mathematically? Perhaps the most important feature of a rooted phylogenetic tree or network is the set of "clusters" that the network represents, as clusters suggest putative monophyletic groups and thus provide hypotheses about the evolutionary relatedness of the taxa under consideration. Hence, in this paper, we treat the calculation of rooted phylogenetic networks in a "cluster-centric" manner and usually interpret rooted phylogenetic networks as representations of sets of clusters.

Clusters and Networks

Exactly which clusters does a rooted phylogenetic network N on X represent? This question has two different answers.

Let N be a rooted phylogenetic network on X. We use the term "hardwired clusters" to refer to the set of all clusters $\mathcal{C}_{\mathrm{hard}}(N)$ that are obtainable from a rooted phylogenetic network N in the following way: each tree edge e in N represents precisely one cluster $\gamma(e)$, which is given by the set of all taxa that appear as labels of nodes below e, that is, all labels of nodes that are descendants of the target node of e.

An alternative way to define the set of clusters represented by N is to use the set of all clusters obtainable from the set of trees $\mathcal{T}(N)$ represented by N. We refer to this as the set $\mathcal{C}_{\mathrm{soft}}(N)$ of clusters represented by N in the "softwired" sense. To obtain these clusters directly from the network N, one must treat the in-edges leading to each reticulation r as a set of alternatives, one of which is "on" if and only if all others are "off." A softwired cluster C is then obtained directly from the network by first deciding, for each reticulation, which reticulation edge is on and which is off, and then collecting all taxa that are reachable below some fixed tree edge e without using any reticulation edge that is off.

To understand the relationship between $\mathcal{C}_{\mathrm{hard}}(N)$ and $\mathcal{C}_{\mathrm{soft}}(N)$, consider figure 9.

Note that given a rooted phylogenetic network N, the set of hardwired clusters of N contains one cluster per tree edge

of N, whereas the number of softwired clusters represented by N is exponential in the number of reticulations contained in N, in the worst case. Note also that given a phylogenetic tree T, it holds that $\mathcal{C}_{\text{hard}}(T) = \mathcal{C}_{\text{soft}}(T)$.

Hardwired Networks

Assume that we are given a set of clusters $\mathcal{C}$ on X. A "cluster network" N for $\mathcal{C}$ is a rooted phylogenetic network that represents the set of clusters on X in the hardwired sense and it can be computed efficiently using the "cluster-popping" algorithm (Huson and Rupp 2008). The number of edges that it contains is, at most, quadratic in the number of given clusters. A cluster network is an abstract phylogenetic network that can be used, for example, to provide a combined visualization of a whole set of rooted phylogenetic trees. Indeed, it has recently been shown (Huson et al. 2011) that a cluster network N that represents all clusters of a given set of trees also contains all the trees themselves, if they are bifurcating; otherwise it contains resolutions of them (for an example, see fig. 10*d*).

However, in practice, the resulting network may sometimes be too large and messy to be of real use. As discussed above for the consensus (super) split networks, one way to address this problem is to represent only those clusters that occur in at least p percent of the input trees, where p is a user-defined parameter. The resulting network will then no longer represent all trees in their full resolution, as some of them will occur only in a contracted form.

Softwired Networks

A number of new methods aim at constructing a rooted phylogenetic network N that represents a set of clusters in the softwired sense, motivated by the assumption that the set of clusters that a network represents is its most important feature, as argued above.

Unfortunately, in general, rooted phylogenetic networks interpreted in the softwired sense are computationally hard to work with. Indeed, even just determining whether a given rooted phylogenetic network N contains some given cluster C on X (in the softwired sense) is NP-complete (Kanj et al. 2008). To avoid these computational problems, we restrict our attention to topologically constrained classes of networks. The concepts of a galled tree (Wang et al. 2000; Gusfield et al. 2003), a galled network (Huson et al. 2009), and a level-k network (Choy et al. 2005) all put constraints on how tangled the undirected cycles in a rooted phylogenetic network may be. The algorithm presented in van Iersel et al. (2010), which aims at computing level-k networks, shows particular promise of becoming a general tool for computing rooted phylogenetic networks from different types of data.

Note that a rooted phylogenetic network that is interpreted in the softwired sense usually requires fewer edges to represent a set of clusters than a hardwired one because individual tree edges can represent more than one cluster. An example of this behavior is shown in figure 10. This implies that a rooted phylogenetic network representing all the clusters of some trees may fail to represent the trees themselves.

All cluster-based methods mentioned in this paper are implemented in the program Dendroscope2 (Huson and Scornavacca 2011).

Hybridization Networks

Assume that we are given a set of taxa X that have evolved under a model of evolution that includes both speciation events and descent with modification, as usual, and, in addition, hybridization events. The evolutionary history of the taxa in X can then be represented by a rooted phylogenetic network N on X where the tree nodes correspond to speciation events and the reticulate nodes correspond to putative hybridization events (Maddison 1997; Linder and Rieseberg 2004). A rooted phylogenetic network that is interpreted in this way is called a hybridization network.

We may attempt to determine such a hybridization network computationally when given two or more gene trees on X, the topologies of which differ significantly and we suspect that these differences are created by hybridization

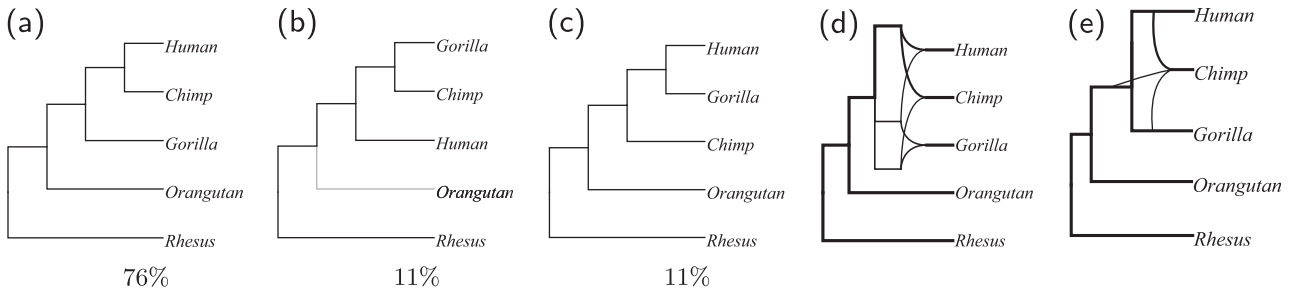

Fig. 10.—Three rooted phylogenetic trees shown in (*a*), (*b*), and (*c*) supported by 76%, 11%, and 11%, respectively, of all genes studied in Ebersberger et al. (2007). In (*d*), we show the cluster network and in (*e*) a (multicombining) galled tree, both representing all clusters contained in the three rooted phylogenetic trees. The line width of each edge is proportional to the number of trees that contain it. Adapted from Huson and Rupp (2008).

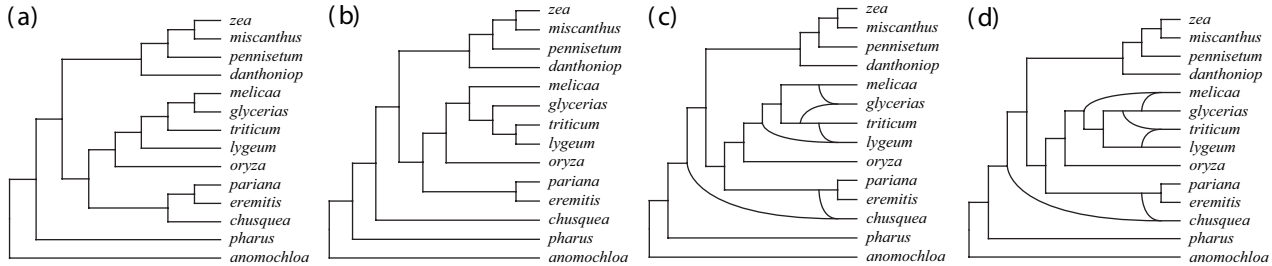

Fig. 11.—Two rooted phylogenetic trees (*a*) T_1 and (*b*) T_2, on 14 grasses, based on the *phyB* gene and *waxy* gene (Grass Phylogeny Working Group 2001). The two rooted phylogenetic networks shown in (*c*) and (*d*) both contain T_1 and T_2, each using a minimum number (three) of reticulate nodes. Each network displays a set of putative hybridization events that may explain the differences between two trees.

events. The corresponding computational problem can be formulated as follows. Given a set T of two or more rooted phylogenetic trees on X, determine a rooted phylogenetic network N that contains all trees in T and has a "minimum" number of reticulate nodes. This is known to be a computationally hard problem (Bordewich and Semple 2007).

Algorithms relevant to this problem, when the input consists of two bifurcating trees, can be found in Baroni et al. (2006); Bordewich et al. (2007); Whidden et al. (2010).

In practice, these algorithms appear to run reasonably fast in many cases. No comparable algorithm exists at present for solving the problem on more than two input trees.

An application is shown in figure 11, where we display two trees, T_1 and T_2, computed for 14 different species of grass (*Poaceae*), based on the *phyB* and *waxy* genes, respectively; see Grass Phylogeny Working Group (2001). Both the two networks shown in figure 11*c* and *d* contain both trees and have the minimum number of reticulate nodes with this property, namely three. If we assume that differences in the topology of the two trees T_1 and T_2 are a result of hybridization events, then, for example, the network in (*c*) suggests that *P. glyceria* is a hybrid of the lineages leading to *P. melica* and *P. triticium*. In the case of the two other putative hybrid species, *P. lygeum* and *P. chusquea*, their evolution requires the postulation of additional lineages to resolve the fact that they appear to be hybrids of recent and less recent lineages. We emphasize that neither network "proves" that hybridization is the cause of the incongruence between trees T_1 and T_2, and additional biological evidence is required to support suspected cases of speciation by hybridization.

Recombination Networks

Assume that we are given a set of taxa X that have evolved under a model of evolution that includes, as usual, both speciation events and descent with modification and also recombination events. The evolutionary history of the taxa in X can then be represented by a rooted phylogenetic network N on X where the tree nodes correspond to speciation events and the reticulate nodes correspond to recombination events. In addition, we require that the following two labelings are given (Griffiths and Marjoram 1996; Gusfield et al. 2003; Huson and Klöpper 2005; Song and Hein 2005):

1. a labeling of all nodes by sequences, and
2. a labeling of all tree edges by the positions in the sequences at which mutations occur.

These labels must be compatible in the sense that the sequences assigned to tree nodes of the network differ exactly by the indicated mutations, whereas the sequences assigned to reticulate nodes must be obtainable from the sequences assigned to the parent nodes by a crossover. A rooted phylogenetic network that is augmented and interpreted in this way is called a recombination network.

An early approach to the problem of computing a recombination network (Hein 1993) is based on the idea of assigning a rooted phylogenetic tree to each position of a given multiple sequence alignment M on X in a most parsimonious way and then combining all the trees into a suitable rooted phylogenetic network N. When doing this, a trade-off must be made between the number of incompatibilities between a character and its associated local tree on the one hand, and, on the other, the "recombination cost" of switching from one tree topology to a different one when going from position i to $i + 1$.

Although the approach has to solve two NP-hard problems and is not practical, it is conceptually appealing because it explicitly addresses the "mosaic" nature of aligned sequences: A long multiple sequence alignment consists of stretches of sequence that have evolved along a common rooted phylogenetic tree, and these stretches are separated by crossover positions at which recombinations have occurred.

More recently, Gusfield et al. (2003) have established a different approach. To obtain a problem that is computationally tractable, they restrict their attention to recombination networks that have the galled tree property. A rooted phylogenetic network N is called a galled tree if every reticulation edge contained in a nontrivial "biconnected component" of N leads to the same reticulation node r. This approach finds a recombination network that is a galled

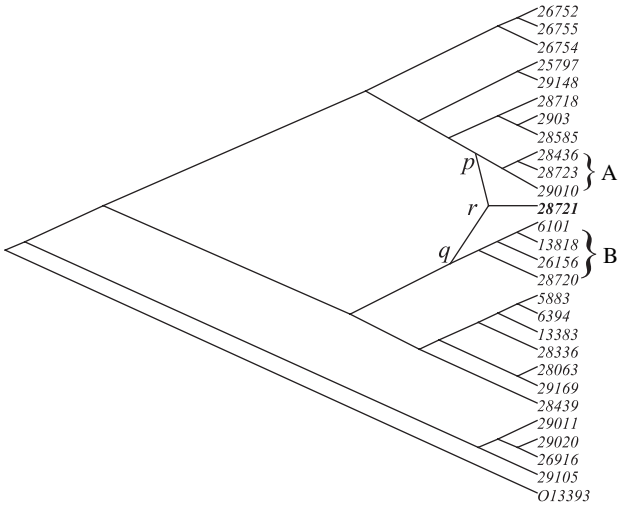

Fig. 12.—A recombination network *N* computed for a multiple sequence alignment of *TRI101* sequences (data from O'Donnell et al. [2000]). This network suggests that the sequence of taxon 28721 arose by recombination of the lineages labeled *A* and *B*. Sequences are not placed at the nodes of the network because of their length.

tree, if any exists, in polynomial time. An application is shown in figure 12.

Gusfield's initial papers on galled trees (Gusfield et al. 2003; Gusfield 2005) generated a lot of interest in this topic. Other papers in this area include Gusfield and Bansal (2005); Huson et al. (2005); Huson and Klöpper (2005); Song (2006); Gusfield et al. (2007). By developing and improving the lower and upper bounds for the number of recombinations required by a data set, Song et al. (2005) have developed a new approach that can be used (in theory) to compute a minimal recombination network. A new approach aimed at a computing a putative recombination history in practice is presented in Parida et al. (2008).

Work on recombination in the context of population genetics by Hein and his colleagues goes far beyond the one approach that we described briefly above. For example, under the "coalescent-with-recombination" model of population genetics, a description of the history of *n*-sampled sequences going backward in time gives rise to a graph that is called an "ancestral recombination graph" (Griffiths and Marjoram 1997; Hein et al. 2005; Song and Hein 2005). This graph is used to perform statistical analyses of the inheritance and prevalence of genes in populations, and the specific topology is often treated as a nuisance variable and integrated out.

DLT Networks

Assume that we are given a set of taxa *X* that have evolved under a model of evolution that includes, as usual, both speciation events and descent with modification as well as gene duplication, loss, and horizontal gene transfer events (Delwiche and Palmer 1996; Planet et al. 2003).

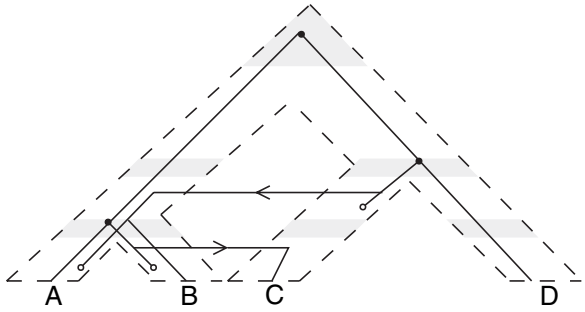

Fig. 13.—An evolution scenario for a gene tree *G* (plain lines) along a species tree *S* (dotted tubes), where the symbol ○ represents a loss. Adapted from Doyon et al. (2010).

The associated computational problem can be formulated as follows: Given a gene tree *T* and a corresponding species tree T_{sp}, reconcile all differences between the two trees by postulating an appropriate DLT scenario. Such a scenario provides a mapping of the gene tree onto the species tree that implies certain duplication, loss, and transfer events. Because the presence of horizontal gene transfer events, this DTL scenario can be seen as a network. An example is shown in figure 13.

Recently, two fast algorithms for a inferring most parsimonious DLT scenario have been proposed. The one described in Tofigh et al. (2010) may propose scenarios that are not time consistent (because of lateral gene transfer events) and considers losses only a posteriori, whereas the other (Doyon et al. 2010) needs the species tree to be dated so as to avoid time-inconsistent scenarios.

Other Types

A number of other types of rooted phylogenetic networks have been developed. We now briefly discuss three of them.

Reassortment Networks

Many viruses are organized into segments of sequence and evolve both by descent with modification and also by reassortment, a process by which viruses that have coinfected a host exchange segments of their genomes. Reassortment is an important mechanism. For example, a possible route to infection of humans by avian strains of influenza A is for swine to be coinfected by avian and human viruses, which reassort to produce a new virus carrying both avian- and human-adapted genes (Castrucci et al. 1993).

A reassortment network is a directed graph in which the nodes represent viral isolates and the edges represent the evolutionary history of the viruses, including reassortment events. Edge weights reflect the edit costs of reassortment and mutation events. Such a graph is organized in layers that correspond to evolutionary stages, such as the seasons in which the viruses were isolated (Bokhari and Janies 2008).

Networks from Multilabeled Trees

Gene duplication is a common event in evolution and so many genes are present in multiple copies in a genome. When some taxa are represented by multiple copies of a gene in a phylogenetic tree, then the tree is called "multilabeled." To analyze the duplication history of a gene, it may be helpful to map such a multilabeled phylogenetic tree onto a single-labeled rooted phylogenetic network so as to see which parts of the tree are similar and which are different (Huber et al. 2006). Algorithms for constructing such a network are discussed in Huson et al. (2011) and are implemented in the program Dendroscope2 (Huson and Scornavacca 2011).

Networks from Rooted Triples

As already mentioned, mathematicians are interested in developing methods that infer a phylogenetic tree or network from basic building blocks. In the computation of a rooted tree or network, these are rooted phylogenetic trees on three taxa, which are sometimes called "rooted triples." In this context, the input is a set R of rooted triples on X, and the goal is to compute a rooted phylogenetic network N that contains all the rooted triples in R and is optimal in some sense. One possible optimality criterion is to minimize the "level" of the network N, which is defined as the maximum number of reticulation nodes contained in any biconnected component of the network in (Jansson et al. 2006). In To and Habib (2009), the authors describe an algorithm that can compute the level-k network with minimum number of reticulations (if such a network exists), for every fixed k in polynomial time.

Conclusions

For unrooted phylogenetic networks, most of the methods mentioned here are routinely used in phylogenetic analysis or phylogeography, particularly Neighbor-Net, consensus split (super) networks and median-joining, given distances, trees, or sequences, respectively.

This is not the case for rooted phylogenetic networks. Although a number of algorithms have been described for computing rooted phylogenetic networks, some problems must be overcome. First, many of the algorithms have only proof-of-concept implementations that are not designed to be used as tools in real studies. Second, the computational problems are often hard, and the algorithms have impractical running times. Third, the calculation of rooted phylogenetic networks must be more closely linked to detailed biological models of reticulate evolution so as to produce more plausible results.

At present, none of the existing methods for computing a rooted phylogenetic method is widely or routinely used as a tool to help understand the evolutionary history of a given set of taxa in terms of mutations, speciations, and specific types of reticulate events. Although rooted phylogenetic networks are conceptually very appealing, the development of suitable methods for their computation remains a formidable challenge.

Literature Cited

Bandelt H-J. 1994. Phylogenetic networks. Verhandlungen des Naturwissenschaftlichen Vereins in Hamburg. 34:51–71.

Bandelt H-J, Dress A. 1993. A relational approach to split decomposition. In: Opitz O, Lausen B, Klar R, editors. Information and classification. Berlin (Germany): Springer. pp. 123–131.

Bandelt H-J, Dress AWM. 1992. A canonical decomposition theory for metrics on a finite set. Adv Math. 92:47–105.

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human population using median networks. Genetics. 141(2):743–753.

Baroni M, Semple C, Steel M. 2006. Hybrids in real time. Syst Biol. 55(1):46–56.

Bokhari S, Janies D. 2008. Reassortment networks for investigating the evolution of segmented viruses. IEEE/ACM Trans Comput Biol Bioinform. 7(2):288–298.

Bordewich M, Linz S, St. John K, Semple C. 2007. A reduction algorithm for computing the hybridization number of two trees. Evol Bioinform. 3:86–98.

Bordewich M, Semple C. 2007. Computing the minimum number of hybridisation events for a consistent evolutionary history. Discrete Appl Math. 155(8):914–928.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21(2):255–265.

Buneman P. 1971. The recovery of trees from measures of dissimilarity. In: Hodson FR, Kendall DG, Tautu P, editors. Mathematics in the archaeological and historical sciences. Edinburgh: Edinburgh University Press. pp. 387–395.

Casado F, Bonvicino CR, Seuanez HN. 2007. Phylogeographic analyses of Callicebus lugens (Platyrrhini, Primates). J Hered. 98(1):88–92.

Castrucci M, et al. 1993. Genetic reassortment between avian and human influenza A viruses in Italian pigs. Virology. 193(1): 503–506.

Choy C, Jansson J, Sadakane K, Sung W-K. 2005. Computing the maximum agreement of phylogenetic networks. Theor. Comput. Sci. 335(1):93–107.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. Mol Ecol. 9:1657–1659.

Delwiche CF, Palmer JD. 1996. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. Mol Biol Evol. 13:873–882.

Dessimoz C, Margadant D, Gonnet GH. 2008. DLIGHT—lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In: Vingron, Martin; Wong, Limsoon, editors. RECOMB 2008. Research in Computational Molecular Biology: Proceedings of the 12th International Conference on Research in Computational Molecular Biology (RECOMB), Volume 4955 of LNCS; Singapore. Berlin (Heidelberg): Springer. pp. 315–330.

Disotell T. 2003. Discovering human history from stomach bacteria. Genome Biol. 4(5):213.

Doolittle WF. 1999. Phylogenetic classification and the Universal Tree. Science. 284:2124–2128.
Doyon J, et al. 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: Tannier, Eric, Editors. RECOMB-CG 2010. Research in Computational Molecular Biology: Proceedings of the 14th International Conference on Research in Computational Molecular Biology (RECOMB), Volume 6398 of LNCS. Ottawa, Canada. Berlin (Heidelberg): Springer. p.*
Dress AWM, Huson DH. 2004. Constructing splits graphs. IEEE/ACM Trans Comput Biol Bioinform. 1(3):109–115.
Ebersberger I, et al. 2007. Mapping human genetic ancestry. Mol Biol Evol. 24(10):2266–2276.
Excoffier L, Smouse P. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics. 136:343–359.
Grass Phylogeny Working Group. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). Ann Mo Bot Gard. 88(3):373–457.
Griffiths RC, Marjoram P. 1996. Ancestral inference from samples of DNA sequences with recombination. J Comput Biol. 3:479–502.
Griffiths RC, Marjoram P. 1997. An ancestral recombination graph. In: Donnelly P, Tavaré S, editors. Progress in population genetics and human evolution, volume 87 of IMA volumes of mathematics and its applications. Berlin (Germany): Springer. p. 257–270.
Grünewald SKF, Dress AWM, Moulton V. 2007. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. Mol Biol Evol. 24(2):532–538.
Gusfield D. 2005. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. J Comput Syst Sci. 70:381–398.
Gusfield D, Bansal V. 2005. A fundamental decomposition theory for phylogenetic networks and incompatible characters. In: Miyano S, et al., editors. RECOMB 2005. Research in Computational Molecular Biology: Proceedings of the Ninth International Conference on Research in Computational Molecular Biology (RECOMB), Volume 3500 of LNCS; Cambridge: Berlin (Heidelberg): Springer. p. 217–232.
Gusfield D, Bansal V, Bafna V, Song YS. 2007. A decomposition theory for phylogenetic networks and incompatible characters. J Comput Biol. 14(10):1247–1272.
Gusfield D, Eddhu S, Langley C. 2003. Efficient reconstruction of phylogenetic networks with constrained recombinations. In: CSB 2003. Proceedings of the IEEE Computer Society Conference on Bioinformatics (CSB); Stanford. Los Alamitos (CA): IEEE Computer Society. p. 363.
Hein J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J Mol Evol. 36:396–405.
Hein J, Schierup MH, Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford: Oxford University Press.
Holland B, Huber K, Moulton V, Lockhart PJ. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol. 21:1459–1461.
Holland B, Moulton V. 2003. Consensus networks: a method for visualizing incompatibilities in collections of trees. In: Benson G, Page R, editors. Benson, Gary, Page, Roderic, editors. WABI 2003; Budapest, Hungary. Heidelberg (Germany). Algorithms in bioinformatics. Proceedings of the Third International Workshop on Algorithms in Bioinformatics (WABI), Volume 2812. Springer. p. 165–176.
Huber KT, Oxelman B, Lott M, Moulton V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. Mol Biol Evol. 23(9):1784–1791.
Huson D, Rupp R. 2008. Summarizing multiple gene trees using cluster networks. Crandall, Keith, Lagergren, Jens, editors. WABI 2008; Budapest, Hungary. Heidelberg (Germany). Algorithms in Bioinformatics: Proceedings of the Eighth International Workshop on Algorithms in Bioinformatics (WABI), Volume 5251 of LNBI. p. 211–225.
Huson D, Scornavacca C. 2011. Dendroscope 2—a program for computing and drawing rooted phylogenetic trees and networks [Internet]. Software Available from: www.dendroscope.org.
Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 14(10):68–73.
Huson DH. 2007. Split networks and reticulate networks. In: Gascuel O, Steel MA, editors. Reconstructing evolution: new mathematical and computational advances. Oxford: Oxford University Press. p. 247–276.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.
Huson DH, Dezulian T, Klöpper T, Steel MA. 2004. Phylogenetic supernetworks from partial trees. IEEE/ACM Trans Comput Biol Bioinform. 1(4):151–158.
Huson DH, Klöpper T, Lockhart PJ, Steel MA. 2005. Reconstruction of reticulate networks from gene trees. In: Miyano S, et al., editors. RECOMB 2005. Research in Computational Molecular Biology: Proceedings of the Ninth International Conference on Research in Computational Molecular Biology (RECOMB), Volume 3500 of LNCS; Cambridge, USA; Berlin (Heidelberg): Springer. p. 233–249.
Huson DH, Klöpper TH. 2005. Computing recombination networks from binary sequences. Bioinformatics. 21(Suppl 2):ii159–ii165.
Huson DH, Rupp R, Berry V, Gambette P, Paul C. 2009. Computing galled networks from real data. Bioinformatics. 25(12):i85–i93.
Huson DH, Rupp R, Scornavacca C. 2011. Phylogenetic networks: concepts, algorithms and applications. Cambridge, UK: Cambridge University Press.
Jansson J, Sung W-K. 2006. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. Theor. Comput. Sci. 363(1):60–68.
Jin G, Nakhleh L, Snir S, Tuller T. 2006a. Efficient parsimony-based methods for phylogenetic network reconstruction. ECCB 2006. Proceedings of the 5th European Conference on Computational Biology (ECCB), Volume 23 of Bioinformatics; Eilat, Israel. Oxford, UK . p. e123–e128.
Jin G, Nakhleh L, Snir S, Tuller T. 2006b. Maximum likelihood of phylogenetic networks. Bioinformatics. 22(21):2604–2611.
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.
Kanj IA, Nakhleh L, Than C, Xia G. 2008. Seeing the trees and their branches in the network is hard. Theor Comput Sci. 401(1–3):153–164.
Kivisild T, et al. 1999. Deep common ancestry of Indian and western Eurasian mtDNA lineages. Curr Biol. 9:1331–1334.
Linder CR, Rieseberg LH. 2004. Reconstructing patterns of reticulate evolution in plants. Am J Bot. 91:1700–1708.
Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46(3):523–536.
Makarenkov V. 2001. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics. 17(7):664–668.
Morrison D. 2005. Networks in phylogenetic analysis: new tools for population biology. Int J Parasitol. 35(5):567–582.

Morrison D. 2010. Phylogenetic networks in systematic biology (and elsewhere). In: Mohan R, editor. Research advances in systematic biology. Trivandrum (India): Global Research Network. p. 1–48.
O'Donnell K, Kistler H, Tacke B, Casper H. 2000. Gene genealogies reveal global phylogeographic structure and reproductive isolation among lineages of *fusarium graminearum*, the fungus causing wheat scab. Proc Natl Acad Sci U S A. 97(14):7905–7910.
Parida L, Melé M, Calafell F, Bertranpetit J. 2008. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. J Comput Biol. 15(9):1133–1153.
Planet PJ, Kachlany SC, Fine DH, DeSalle R, Figurski DH. 2003. The widespread colonization island of Actinobacillus actinomycetemcomitans. Nat Genet. 34:193–198.
Rieseberg LH. 1997. Hybrid origins of plant species. Annu Rev Ecol Evol Syst. 28:359–389.
Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.
Sneath P. 1975. Cladistic representation of reticulate evolution. Syst Zool. 24(3):360–368.
Song YS. 2006. A concise necessary and sufficient condition for the existence of a galled-tree. IEEE/ACM Trans Comput Biol Bioinform. 3(2):186–191.
Song YS, Hein J. 2005. Constructing minimal ancestral recombination graphs. J Comput Biol. 12(2):147–169.
Song YS, Wu Y, Gusfield D. 2005. Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution. Bioinformatics. 21:413–422.
Syvanen M. 1985. Cross-species gene transfer; implications for a new theory of evolution. J Theor Biol. 112(2):333–343.
Templeton AR, Crandall KA, Sing CF. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics. 132:619–633.
To T-H, Habib M. 2009. Level-k phylogenetic networks are constructable from a dense triplet set in polynomial time. Combinatorial Pattern Matching. In: Kucherov, Gregory, Ukkonen, Esko, editors. CPM 2009. Proceeding of the 20th Annual Symposium Combinatorial Pattern Matching (CPM), Volume 5577 of LNCS Lille, France: Heidelberg (Germany). p. 275–288.
Tofigh A, Hallet M, Lagergren J. 2010. Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans Comput Biol Bioinform. 99.
van Iersel L, Kelk S, Rupp R, Huson D. 2010. Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. ISMB 2010. Proceedings of the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Volume 26 of Bioinformatics; Global Boston, USA. Oxford (UK). p. i124–i131.
Wagele J, Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol Biol. 7(1):147.
Wang L, Ma B, Li M. 2000. Fixed topology alignment with recombination. Discrete Applied Mathematics. 104(1–3):281–300.
Whidden C, Beiko R, Zeh N. 2010. Fast FPT algorithms for computing rooted agreement forests: theory and experiments. In: Festa, Paola, editors. SEA 2010. Proceedings of the 9th International Symposium on Experimental Algorithms (SEA), Volume 6049 of LNCS. Naples, Italy: Heidelberg (Germany). p. 141–153.
Whitfield J, Cameron S, Huson D, Steel M. 2008. Filtered Z-closure supernetworks for extracting and visualizing recurrent signal from incongruent gene trees. Syst Biol. 57(6):939–947.

Associate editor: Bill Martin

Genome Biology and Evolution SMBE