MBE Advance Access published October 12, 2005

Research Article (Submitted to MBE 2005)

Application of Phylogenetic Networks in Evolutionary Studies

Daniel H. Huson
Center for Bioinformatics (ZBIT),
Tübingen University, Sand 14, 72076 Tübingen, Germany,
Phone: +49 7071 2970450, Fax: +49 7071 295148,
Email: huson@informatik.uni-tuebingen.de

David Bryant
Department of Mathematics Auckland University
Phone: +64 9 373 7599 x 88763, Fax: +64 9 373 7457
Email: bryant@math.auckland.ac.nz

September 26, 2005

Abstract

The evolutionary history of a set of taxa is usually represented by a phylogenetic tree, and this model has greatly facilitated the discussion and testing of hypotheses. However, it is well known that more complex evolutionary scenarios are poorly described by such models. Further, even when evolution proceeds in a tree-like manner, analysis of the data may not be best served by using methods that enforce a tree structure, but rather by a richer visualization of the data to evaluate its properties, at least as an essential first step. Thus, phylogenetic networks should be employed when reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and -loss are believed to be involved, and, even in the absence of such events, phylogenetic networks have a useful role to play. This paper reviews the terminology used for phylogenetic networks and covers both split networks and reticulate networks, how they are defined and how they can be interpreted. Additionally, the paper outlines the beginnings of a comprehensive statistical framework for applying split network methods. We show how split networks can represent confidence sets of trees and introduce a conservative statistical test for whether the conflicting signal in a network is treelike. Finally, this paper describes a new program SplitsTree4, an interactive and comprehensive tool for inferring different types of phylogenetic networks from sequences, distances and trees.

Keywords: phylogeny, networks, software, confidence intervals

Running head: Phylogenetic Networks

Introduction

The familiar evolutionary model assumes a tree, a model that has greatly facilitated the discussion and testing of hypotheses. However, it is well known that more complex evolutionary scenarios are poorly described by such models. Even when evolution proceeds in a tree-like manner, analysis of the data may not be best served by forcing the data onto a tree, or assuming a tree-like model. Rather, visualization and exploration of the data to discover and evaluate its properties can be an essential first step. Recognition of theses issues has led to the development of a number of different types of phylogenetic networks, see Figure 1.

[Figure 1 about here.]

In this article we first discuss the terminology used to describe and distinguish between the different types of phylogenetic networks and methods that are currently in use. There are a number of different types of phylogenetic networks and this can be a source of confusion.

We then provide a detailed introduction to split networks, how they are defined and how they can be interpreted. We discuss the use of split networks in strategies for dealing with systematic error, especially as a preliminary step to tree-based analysis, and show examples of a number of different types of networks.

We outline the beginnings of a statistical framework for phylogenetic inference using split network methods. Split networks have been widely used as a tool for visualization, but have been neglected as a tool for statistical inference. Their success as a tool for visualizing incompatible and ambiguous phylogenetic signals suggests that, properly used, they can become an invaluable statistical inference tool for phylogenetics. Here we describe a method for constructing approximate confidence intervals for trees using networks, and a simple test to determine whether incompatible signal in a network is statistically significant. There is considerable scope for improvement of both methods.

Finally, we present our new program SplitsTree4, which provides a comprehensive and interactive frame-work for estimating phylogenetic trees and networks. The program provides methods for computing split networks *from sequences*, such as the median network (Bandelt et al., 1995) and networks based on spectral analysis (Hendy and Penny, 1993); *from distances*, such as split decomposition (Bandelt and Dress, 1992) and neighbor-net (Bryant and Moulton, 2004); and *from trees*, such as consensus networks (Holland et al., 2004) and super networks (Huson et al., 2004). Moreover, new methods are provided for computing hybridization networks from trees (Huson et al., 2005) and recombination networks from binary sequences (Huson and Kloepper, 2005). Additionally, SplitsTree4 includes a large number of different distance computations and tree building methods.

Terminology

A *phylogenetic tree* is commonly defined as a leaf-labeled tree that represents the evolutionary history of a set of taxa, possibly with branch lengths, either unrooted or rooted.

The concept of a “phylogenetic network” is much less well defined and there exist many different usages of the term, see Figure 1. We propose to define a *phylogenetic network* as *any* network in which taxa are represented by nodes and their evolutionary relationships are represented by edges. (For phylogenetic trees, edges are referred to as branches.) Under this very general heading, one can distinguish between a number of different types of networks. *Phylogenetic trees* are one type, see Figure 2. A second type are *split networks*, which are obtained as a combinatorial generalization of phylogenetic trees, and are designed to represent incompatibilities within and between data sets, see Figure 3(a). A third type, *reticulate networks*, represent evolutionary histories in the presence of reticulate events such as hybridization, horizontal gene transfer or recombination, see Figure 3(b). A number of additional types of networks exist (Posada and Crandall, 2001), for example, to represent gene duplication and loss phylogenies (Hallett and Lagergren, 2000; Durand et al., 2005) or host and parasite co-evolution (Charleston, 1998). Other approaches for constructing phylogenetic networks include *statistical parsimony* (Templeton et al., 1992), the *netting* method (Fitch, 1997), and a method that, in effect, consists of adding “short-cut” edges to a tree (Legendre and Makarenkov, 2002).

One major source of confusion has been that different authors define the generic term “phylogenetic network” rather narrowly to mean some particular type of network currently under study. For example, a recent paper on recombination (Gusfield and Bansal, 2005) defines a phylogenetic network to be a “recombination network”, whereas a recent paper on hybridization (Linder and Rieseberg, 2004), defines a phylogenetic network to be a “hybridization network".

[Figure 2 about here.]

[Figure 3 about here.]

Reticulate networks provide an *explicit* representation of evolutionary history, generally depicted as a phylogenetic tree with additional edges. The internal nodes in such a network represent ancestral species and nodes with more than two parents correspond to reticulate events such as hybridization or recombination.

Split networks are used to represent incompatible and ambiguous signals in a data set. In such a network, parallel edges, rather than single branches, are used to represent the splits computed from the data. To be able to accommodate incompatible splits, it is often necessary that a split network contains nodes that do not represent ancestral species. Thus, split networks provide only an *implicit* representation of evolutionary history.

The distinction between these explicit and implicit representations of evolution is important and has not always been made clear in the past (Morrison, 2005).

Background

A *split* is a partition of the taxa into two non-empty subsets, such as the partition obtained when we remove a branch from a phylogenetic tree. For example, removing the branch indicated by the arrow in Figure 2(a) splits the taxa into two groups $\{B, C, D, E\}$ and $\{o, A, F, G, H, I, J, K\}$. The interpretation of split networks is based on one simple principle: *a split network contains exactly the same information as a list of splits with a weight for each split.*

In a split network, every edge is associated with a split of the taxa, but there may be a number of parallel edges associated with each split. In Figure 4, all of the edges corresponding to a split are highlighted, as well as all the taxa on one side of the split. If the edges associated with a particular split were to be deleted, then the network would become disconnected with precisely two components, corresponding to the two parts of the split. The edges separate taxa on one side of the split from the taxa on the other side of the split. The length of an edge in the network is proportional to the weight of the associated split. This is analogous to the length of a branch in a phylogenetic tree.

The example in Figure 4 gives two representations of the same information: (a) a split network representing 14 different splits of a set of 8 taxa; and (b), a listing of the splits and their weights.

[Figure 4 about here.]

Formally, for a given taxon set X and set of splits $\mathcal{S}$, we define a *split network* $\mathcal{N}$ to be a connected graph in which some of the nodes are labeled by taxa and all edges are labeled by splits, such that:

(N1) Removing all edges associated with a given split S in $\mathcal{S}$ divides $\mathcal{N}$ into two connected components, one part containing all taxa on one side of S and the other part containing all taxa on the other side.

(N2) The edges along any shortest path in $\mathcal{N}$ are all associated with different splits.

Every split network represents a unique collection of splits. However, uniqueness does not hold in the other direction, as a given collection of splits can have many different split network representations. In Figure 5 we show two different split networks that both represent the splits $ABC|DEF$, $ABF|CDE$, $AEF|BCD$, together with all six trivial splits on the set $\{A, B, C, D, E, F\}$ (Wetzel, 1995).

[Figure 5 about here.]

Due to this non-uniqueness, it is often inappropriate to consider internal nodes as hypothetical ancestors (Bandelt and Dress, 1992). The interpretation of split networks and statistical tests on them, are based on the underlying set of splits. One exception is the use of split networks to visualize distance matrices. The *phenetic distance* between two taxa in a split network is defined as the sum of the weights (or lengths) of the edges along a shortest path between the taxa (Bryant and Moulton, 2004). This distance can be computed directly from the associated splits and weights, and does not change for different split network representations.

The split network, then, is a graphical representation of a collection of splits with weights. The interpretation of the network therefore depends on exactly how the splits were constructed and assigned weights. As we shall see, this varies considerably between methods and between applications.

Interpreting split networks: representing multiple trees

There are many situations in phylogenetics where we have a large collection of trees that we wish to summarize in some way. The trees might be the result of a bootstrap analysis, samples from a posterior distribution or come from a multi-gene analysis, for example. Techniques for summarizing multiple trees using split networks are described in (Bandelt, 1995; Holland et al., 2004; Huson et al., 2004). The basic idea is to code each individual tree as a collection of splits, define a summary set of splits from these and represent the resulting set using a split network. For example, *consensus networks* are constructed from all splits appearing in at least some fixed proportion of the input trees. A consensus network can represent much more information than a single tree with p-values.

Here we take the consensus network methods one step further and use split networks to define *confidence sets* of trees. Suppose that we have assigned an interval for the weights of each split (represented) in a split network $\mathcal{N}$. We say that a tree T is contained within the split network $\mathcal{N}$ if

1. Every split in the tree is a split in the network.

2. For every split in the tree, the corresponding branch length is contained within the interval assigned to the appropriate split weight.

3. For every split in the network *not* in the tree, the interval assigned to that split contains zero.

For example, the split network in Figure 6(a) contains the tree T_1 but not the tree T_2, since 0 is not in the interval assigned to the split $AB|CD$.

[Figure 6 about here.]

A geometric interpretation might prove useful here. Suppose that the splits are indexed from 1 to m. A tree can then be coded as a point in m-dimensional space: the ith co-ordinate is the length corresponding to the ith split, or 0 if that split is not present in the tree (Holmes, 2005). The split network then corresponds to a box in m-dimensional space: the range of values in the ith dimension is given by the interval for the ith split. A tree is contained in the network if the corresponding point is contained in the box.

A network with intervals assigned to edges is an *X% confidence network* if, for different random samples, it has an X% probability of containing the 'true' tree. In the appendix we propose a method for constructing *approximate* confidence networks using non-parametric bootstrapping. Confidence networks can also be defined for Bayesian analyses, where they would describe posterior confidence sets. The split summaries produced by MrBayes (Ronquist and Huelsenbeck, 2003) come close to this.

The use of split networks to describe confidence sets for trees fits well with the geometric analysis of tree-space in (Billera et al., 2001). They construct a model of tree-space by pasting together sections of Euclidean space. Each dimension, in their picture, corresponds to a different split. Hence their model of tree-space sits naturally in the much larger, and much simpler, space of split networks.

Interpreting split networks: networks and systematic error

The rapid growth in available genomic sequence data opens up exciting new possibilities, but also new challenges, for phylogenetic inference. This development means that *sampling error* is becoming less of an issue while the impact of *systematic error* is becoming increasingly important. Sampling error is random error resulting from a small sample size (number of sites). Systematic error occurs when mistakes in the assumptions of a model or method cause data to be mis-interpreted, something that is even more likely to occur when we consider large, multi-gene, heterogeneous data sets. These cause biases and artefacts in phylogenetic inference, some of which can be corrected by modifying the employed model of sequence evolution (Delsuc et al., 2005; Bryant et al., 2005; Swofford et al., 1996; Felsenstein, 2004a). Systematic error is particularly important when there is a possibility of reticulate evolutionary events, since, in such cases, no tree-based model can accurately model the data.

The two most widely used methods for checking the reliability of a tree are the non-parametric bootstrap (Felsenstein, 1985) and (for Bayesian analysis) multiple samples from the posterior distribution (Rannala and Yang, 1996; Ronquist and Huelsenbeck, 2003). These techniques are designed to protect from sampling error; they are *not* designed to protect from systematic error. On short sequences, bootstrap resampling has the effect of 'jiggling' the data and so can provide some assessment of robustness. However, when the number of sites is very large, the bootstrap replicates will be all very similar and so the bootstrap support will be high, no matter how poorly the data fits a tree.

Systematic error, unlike sampling error, does not disappear as the sequence length increases. Indeed it might even become worse (Delsuc et al., 2005).

In tree-based phylogenetic analysis the goal is to find the phylogenetic tree that best explains the observed patterns in the data. When there is systematic error, shortcomings of the model mean that the observed data will (in general) not appear to have originated from any tree. The tree-building method will attempt to fit a tree, even if there is still a huge gap between the data and the best tree that the method can find. This gap is the principal cause of reconstruction artefacts (Steel, 2005).

Model-based split network methods (e.g. Bryant and Moulton (2004); Huber et al. (2002); Winkworth et al. (2005)) deal with systematic error by adding parameters to the evolutionary model. In conventional phylogenetics, parameters are added to give a more complex model of the substitution process down a branch. However, there are two kinds of parameters in phylogenetic inference: parameters describing the evolutionary model (rate variation, substitution probabilities, etc.) and parameters describing the topology (the tree and branch lengths). Evolutionary models based on split networks add extra topology-related parameters. A phylogenetic tree corresponds to a collection of compatible splits (with weights or lengths). A split network model is obtained by allowing additional splits with weights. These extra parameters allow split networks to fit the data better than individual trees.

It may seem odd to use split networks, which are not trees, to represent phylogenetic signals that, for the most part, originate from trees. However this is not an unusual practice in statistics. Consider the statement “in the year 2000, the expected number of children in a randomly chosen family in the US was 1.86”. Of course, there was not a single family with exactly 1.86 children, and it does not make sense to talk about a fractional number of children in a given family. It does make sense, however, to use fractions in summary statistics like this one. The same applies to split networks. In the absence of reticulation, it doesn’t make sense to speak of sequences evolving on a network but it does make sense to infer split networks as summary statistics.

The example in (Kolaczkowski and Thornton, 2004) demonstrates that unaccounted rate variation can mislead both maximum likelihood (ML) and parsimony-based analyses into selecting the wrong tree. We show below that the rate variation did not mislead a split network method, split decomposition, which displayed, simultaneously, support for the true tree and the artefactual tree. The split network better represented the phylogenetic signal than either tree.

The example in (Esser et al., 2004) shows how split network methods can extract phylogenetic signals that are missed by tree-based methods. The study involves multiple genes and an ML tree analysis of each gene gives statistical (i.e. bootstrap) support for conflicting phylogenies. However it is apparent that the ML analyses are affected by systematic error. A neighbor-net analysis returns split networks for each gene that incorporate both the ML tree and additional splits. Furthermore, many of these additional splits appear in almost all networks for the other genes. The probability of this occurring by chance is extremely small. Hence the additional splits recovered by the split network method are probably contained in the true underlying phylogeny.

To summarize, an initial strategy for using split network methods in phylogenetic inference would be:

1. Construct a split network using the best available model and method.
2. Determine if the network is significantly different from a tree.
3. If the network is significantly non-treelike then there is probably an error in the model. If possible, improve the model and go back to step 1. If the conflicting or ambiguous signal in the split network can not be explained then this failure to explain the data properly should be reported and taken into account in any conclusions drawn from the phylogenetic analysis.
4. If the network is not significantly non-treelike (and there is evidence that the sampling error of the network method is not too large) then continue with a tree-based phylogenetic analysis.

Reticulate networks

Split networks provide an *implicit* picture of evolutionary relationships. Sets of parallel edges are employed to represent splits. As mentioned above, internal nodes in a split network do not necessarily correspond to hypothetical ancestors.

In contrast, “reticulate networks” provide an *explicit* picture of evolution. In such a network, edges represent lineages of descent or reticulate events such as hybridization, horizontal gene transfer or recombination, and all nodes correspond to hypothetical ancestors, whether the product of speciation and mutation, or reticulate events.

Such explicit networks are usually drawn “rooted”, so that the edges have a direction with an evolutionary meaning. In contrast, most split networks published to date are displayed as unrooted networks. However, it is possible to root split networks, e.g. by specifying an outgroup, as illustrated in Figure 2(a), and an algorithm for drawing rooted split networks is available in SplitsTree4.

A *hybridization network* is a reticulate network $\mathcal{N}$ that can explain a given set of trees in terms of hybridization, see (Maddison, 1997; Baroni et al., 2004; Nakhleh et al., 2004; Huson et al., 2005). More precisely, given a set of trees $\mathcal{T} = \{T_1, \ldots, T_m\}$, usually obtained from a collection of different genes, one would like to determine a putative reticulate network $\mathcal{N}$ from which the trees arise. If such a network can be found, then we say that the network $\mathcal{N}$ *explains* the set of trees $\mathcal{T}$ in terms of hybridization.

In a recent paper (Huson et al., 2005), we describe a method that can solve this problem for certain patterns of hybridization and this is implemented in the SplitsTree4 program. It takes as input a set of trees and first computes the splits network that represents all splits present in the input trees. Each “netted component” (“2-connected component”, in terms of graph-theory) of the network is then individually analyzed in turn and is replaced by a reticulation scenario, if one can be found that explains the given splits in the component.

As an example, suppose that the two trees depicted Figure 2 are based on two different genes. In Figure 3(a) we show the split network that represents all splits present in either of the two trees. The netted regions in the network indicate in which parts of the phylogeny there is disagreement between the two trees.

Figure 3(b) depicts a reticulate network that can explain the differences in the two trees using three reticulation events. In this example, the clade $\{B, C\}$ arises from a reticulation event between the lineages leading to taxa A and D, whereas the taxon H, or I, arises from a reticulation event between the lineages leading to G and $\{J, K\}$, or to $\{F, G\}$ and J, respectively.

Recombination as a reticulate event is usually considered in population studies (Hudson, 1983; Hein, 1990) and the arising graphs are called *ancestor recombination graphs* (ARGs), or *recombination networks.* A number of algorithms for inferring such networks have recently be proposed (Gusfield and Bansal, 2005; Lyngsoe et al., 2005; Huson and Kloepper, 2005). Such methods take as input a sequence of two-state characters and attempt to explain the given characters in terms of evolution by speciation, mutation and recombination, under the restriction that a mutation of any character may happen at most once throughout the whole phylogeny. In (Huson and Kloepper, 2005), we describe a new method that solves this problem for certain patterns of recombination and this approach is implemented in SplitsTree4.

Methods

SplitsTree4 is a completely new program that we have developed over the past three years. It was inspired by SplitsTree3 (Huson, 1998), which was primarily an implementation of the split decomposition method (Bandelt and Dress, 1992). SplitsTree4 integrates a wide range of phylogenetic network and phylogenetic tree methods, inference tools, data management utilities, and validation methods. The key design goals were ease of use, portability, and flexibility. A user can click their way through a split network analysis or control the entire program from a command line. The code is written in Java and thus runs under Linux, MacOS and Windows. The support of plugins makes it easy to add new functionality. The program is freely available from: `www.splitstree.org`.

There are now many published methods for inferring split networks and we have implemented most of them in SplitsTree4. These include

- median networks (Bandelt et al., 1995), parsimony splits (Bandelt and Dress, 1994) and spectral analysis (Steel et al., 1992), which construct split networks (or equivalently, weighted splits) directly from character data;
- split decomposition (Bandelt and Dress, 1992) and neighbor-net (Bryant and Moulton, 2004) which construct split networks from inferred distance matrices;
- consensus networks (Bandelt, 1995; Holland et al., 2004) and super-networks (Huson et al., 2004) which construct split networks from sets

of trees.

SplitsTree4 also contains methods for constructing other kinds of phylogenetic networks, including recombination networks (Huson et al., 2005) and hybridization networks (Huson and Kloepper, 2005).

The software implements maximum likelihood (ML) estimation of distances from amino acid and nucleotide sequences, under all standard evolutionary models. The implemented phylogenetic tree methods include *neighbor-joining* (Saitou and Nei, 1987), *Bio-NJ* (Gascuel, 1997), *UPGMA* (Sokal and Michener, 1958), *the Buneman tree* (Buneman, 1971), the *refined Buneman tree* (Brodal et al., 2003) and standard consensus tree methods. SplitsTree4 also provides a graphical front-end for ML analysis using PhyML (Guindon and Gascuel, 2003) and parsimony analysis using Phylip (Felsenstein, 2004b). All character-based analyses can be bootstrapped.

The graphical interface makes it easy for users to interactively explore their data using different phylogenetic methods, starting with a standard file format (e.g. NEXUS, FastA, Phylip,...). Users can select from different analyses, filter data, and manipulate networks, all using standard menus. The data, and networks, can be exported to a variety of different file formats including the standard image formats PostScript, JPEG, SVG, PNG and GIF. Multiple analyses can be run at the same time, and the program is multi-threaded so time-consuming calculations can be done in the background.

More details on the architecture of the program can be found in the appendix.

Examples

Heterogeneous evolution and split networks

The variation of evolutionary rates across sites and between lineages is a well-recognized source of phylogenetic error. A particularly intriguing synthetic example is described in (Kolaczkowski and Thornton, 2004), intriguing because in their experiments, parsimony outperforms likelihood in situations of model violation (see also Spencer et al. (2005); Steel (2005)). The experiment provides a clear and simple illustration of how violation of the evolutionary model can mislead phylogenetic inference. In effect, sequences that evolved on one tree appear to have evolved on a different tree.

Sequences of varying lengths were evolved on a single tree, half of the sites with one set of branch lengths and the remaining sites with a second set of branch lengths, see Figure 7(a). The change in branch lengths simulates an extreme example of rate variation, one that violates the basic assumptions of standard maximum likelihood analysis. Consequently, ML repeatedly reconstructs the wrong tree, even with long (or indeed, infinite) sequences. Parsimony was also misled, but, in this setup, less so than ML.

We repeated the experiment of (Kolaczkowski and Thornton, 2004) to examine how split decomposition is affected by the model variation. For each length (r) of the internal branch we constructed sequences of length 1000, 10 000, and 100 000. These were analyzed using maximum parsimony, maximum likelihood (using PhyML, Guindon and Gascuel (2003), with the Jukes-Cantor model, no site rate distribution and no invariant sites) and split decomposition. A tree, or network, was judged ‘correct’ if it contained the true split separating A and B from C and D, and at most one alternative. We performed 200 replicates for each parameter setting, and report the proportion correct in Figure 7(b).

Split decomposition returned the ‘correct’ split for even small values of r. Of course, split decomposition gets an unfair advantage: effectively, it can choose two trees instead of one. This is exactly the advantage of networks. In this case, split decomposition cannot decide between two trees, the true tree and the ML tree. The networks returned for different values of r (and infinite sequences) are given in Figure 7(c). Looking, for example, at the split network produced when $r = 0.3$, we see that there is an even balance between support for the tree $AB|CD$ and the tree $AD|BC$. The indecision in the ML analysis is due to sampling error pushing the signal towards one tree or the other. As r increases, the phylogenetic tree methods settle uniquely on a single tree which, in this case, is the correct tree. However the network still has a large box: the closest tree may be the correct tree but there is a substantial amount of phylogenetic signal in the data that is not being explained adequately by a single tree.

[Figure 7 about here.]

Animal phylogeny

Our second case study illustrates the application of SplitsTree4 to genome-scale phylogenetics. The data set was prepared by (Philippe et al., 2005) and consists of 71 (slowly evolving) genes (20,705 amino acid positions) from 35 animals, 10 fungi and 2 choanoflagellates. The phylogeny for the bilaterian animals has been the subject of much discussion recently, one key question being whether the ecdysozoa (e.g. arthropods, nemotodes, tardigrades) or the coelomates (e.g. mollusca, deuterosomes, arthropoda) are monophyletic (Hedges, 2002; Philippe et al., 2005; Wolf et al., 2004).

We conducted a neighbor-net analysis using ML distances inferred from a concatenated data set under the Jones, Taylor, and Thornton (JTT) + F + Γ model (Jones et al., 1992), see Figure 8. One hundred bootstrap replicates were performed. The nemotodes, arthropods, deuterotomes, choanoflagellates and platyhelminthes are all well supported groups with 100% bootstrap support. However the relationships between these groups are less clear: the network seems mid-way between a tree grouping nemotodes, tardigrades and arthropods (the ecdysozoa hypothesis) and a tree grouping arthropods, tardigrades and deuterostomes (the coelomate hypothesis). Neither hypothesis is supported by a clear split in the network: the clearest split in favor of the ecdysozoa hypothesis misplaces the annelids, while the clearest split in favor of the coelomates misplaces the cnidaria.

[Figure 8 about here.]

Neighbor-net, like any phylogenetic method, is affected by sampling error, and this could potentially explain the observed conflicting signals. There are many potential sources for systematic error in a data set this large and diverse (Philippe et al., 2005): rate variation, heterotachy, variation in substitution rates, or interdependence of sites, to name a few. One suspected consequence of this modeling error is an increased problem with long branch attraction, where biases in the model or method tend towards trees grouping long branches together (Felsenstein, 2004a).

Using a subset of the taxa, it was recently demonstrated (Philippe et al., 2005) that support for a coelomates clade or an ecdysozoa clade with a tree-based method depends greatly on the choice of outgroup, see Figure 9. The ecdysozoa clade appears only once a closely related outgroup (a cniderian) was used, indicating that the coelomate tree is due to long branch attraction. This is an example of the effect of taxon instability: the information present in the cniderian sequence changes the resolution of taxa in other parts of the tree. If we look at a split network analysis of the sequences we see that there is at least some support for the ecdysozoa clade even when the cnidarian taxon is absent.

When the cniderian taxon is absent the tree-based method (in this case, BioNJ Gascuel (1997)) consistently ignores that information supporting the ecdysozoa hypothesis, even though the split network (computed by neighbor-net) indicates that there is some support for an ecdysozoa clade. When the cniderian is included, the tree based method switches to consistently supporting the ecdysozoa clade. The support for the ecdysozoa hypothesis does not rest on a single taxon: the additional taxon merely tips the balance. The split network, in contrast, does not change abruptly with the inclusion of the additional taxon (not shown). Both trees are contained in a single network, and the support for one or the other only alters the branch lengths/split weights. (However, in other situations, split networks can also change abruptly after the addition of taxa.)

[Figure 9 about here.]

The evolutionary history of dusky dolphins

This example is taken from a recent paper (Cassens et al., 2003) that investigates the phylogeography of dusky dolphins (*Lagenorhhynchus oscurus*) and compares a number of different network methods. The data used are the sixty variable positions in the DNA sequences of the full mitochondrial cytochrome *b* gene for 36 different haplotypes seen in 124 individuals, sampled off Peru, Argentina and Southwest Africa. The paper discusses the application of four different network methods, split decomposition, the “minimum spanning network” “statistical parsimony”, and the “median joining network”.

We have reanalyzed this data using methods available in SplitsTree4. Calculation of observed-P distances and application of the neighbor-joining method produced the tree depicted in Figure 10(a). The edges are labeled by the %-bootstrap support attained in 1000 bootstrap replicates.

Based on this tree, the 95%-confidence network for neighbor-joining displayed in Figure 10(b) shows that there is considerable sampling variance in the tree estimate, particularly when we consider the large confidence intervals on the edge lengths. One split S' places the Atlantic haplotype $A9$ together with two Pacific haplotypes $P4.1$ and $P4.2$, which is incompatible with the central split S between Pacific and Atlantic haplotypes. The confidence intervals for both splits include 0, so there is too much sampling error with neighbor-joining to discriminate between the two possibilities.

Application of the maximum parsimony algorithm using the Phylip `dnapars` program produced *three* most parsimonious trees, whose consensus network is depicted in Figure 10(c). Here we see that the three trees do not differ significantly.

In Figure 10(d) we display the median network. Each split is labeled by the columns of the alignment that support it. We see that the central split S separating Pacific and Atlantic haplotypes is supported by precisely three positions, $3, 18, 25$, whereas the incompatible split S' is only supported by position 11.

Application of split decomposition to the uncorrected-P distances produces the split network shown in Figure 10(e). On this data set, the split decomposition captures most of the incompatible signals that are present in the median network, including the two incompatible splits mentioned above. Finally, we display the split network produced by neighbor-net in Figure 10(f). By definition, this method can only produce a planar network. In this example the resulting network is missing some of the splits that are present in the median network and also in the split decomposition network. (In theory, and for small or highly similar data sets, the split decomposition method produces more resolved networks than neighbor-net, as the networks produced by split decomposition are not restricted to be planar. However, in practice, and for large or divergent data sets, split decomposition suffers from low resolution and the networks produced by neighbor-net are usually more resolved.)

[Figure 10 about here.]

Discussion

Phylogenetic networks have an important role to play in the reconstruction of evolutionary history. Implicit models such as split networks are very useful for exploring and visualizing the different signals in a data set. Explicit models such as hybridization and recombination networks can be used to provide an explicit description of reticulate evolution. Both types of networks have an important role to play. Many current methods for computing split networks from characters, distances or trees are very robust and can provide valuable insights. In contast, the existing methods for computing hybridization and recombination networks are unproven, and, as all are based on some kind of combinatorial analysis of a given configuration of splits, they are very susceptible to false positive signals.

In the past, split network methods have been neglected as a tool for statistical inference in phylogenetics. This was due, in part, to the lack of an appropriate statistical framework, the absence of an integrated software package, and conceptual difficulties of thinking in terms of split networks as well as trees.

We have described the beginnings of a comprehensive statistical framework for phylogenetic analysis based on split network methods. The key messages are that split networks are representations of splits, and using more splits permits a more accurate representation of the data. The confidence networks and tests we presented illustrate a new general approach, although much remains to be done to improve the efficiency and power of these methods.

Finally, we have introduced the new SplitsTree4 program, which was inspired by the popular SplitsTree3 program. SplitsTree4 is an integrated and user-friendly software package allowing users to conduct phylogenetic analysis using trees, split networks and reticulate networks. With SplitsTree4, we hope to provide a robust framework for inferring and investigating phylogenetic networks.

Acknowledgments

We would like to thank Andreas W.M. Dress for introducing us to splits and phylogenetic networks. We would also like to thank Pete J. Lockhart, Barry G. Hall and Susan Holmes for helpful discussions.

References

Bandelt, H. J. 1995. Combination of data in phylogenetic analysis. Plant Syst. Evol. Suppl. **9**:355–361.

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

———. 1994. A relational approach to split decomposition. Tech. rep., Universität Bielefeld.

Bandelt, H. J., P. Forster, B. C. Sykes and M. B. Richards. 1995. Mitochondrial portraits of human population using median networks. Genetics **141**:743–753.

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

Beran, R. 1988. Balanced simultaneous confidence sets. J. Amer. Statist. Assoc. **83**:679–686.

———. 1990. Refining bootstrap simultaneous confidence sets. J. Amer. Statist. Assoc. **85**:417–426.

Billera, L., S. Holmes and K. Vogtman. 2001. Geometry of the space of phylogenetic trees. Adv. Appl. Math. **27**:733–767.

Brodal, G. S., R. Fagerberg, A. Östlin, C. N. S. Pedersen and S. S. Rao. 2003. Computing refined buneman trees in cubic time. Lect. Notes Comput. Sc. **2812**:259–270.

Bryant, D., N. Galtier and M.-A. Poursat. 2005. Likelihood calculation in molecular phylogenetics. In O. Gascuel, ed., Mathematics of Evolution and Phylogeny, 33–62. Oxford University Press.

Bryant, D. and V. Moulton. 2004. NeighborNet: An agglomerative algorithm for the construction of planar phylogenetic networks. Mol. Biol. Evol. **21**:255–265.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall and P. Tautu, eds., Mathematics in the Archaeological and Historical Sciences, 387–395. Edinburgh University Press, Edinburgh.

Cassens, I., K. van Waerebeek, P. B. Best, E. A. Crespo, J. Reyes and M. C. Milinkovitch. 2003. The phylogeography of dusky dolphins (*lagenorhynchus obscurus*): a critical examination of network methods and rooting procedures. Mol. Ecol. **12**:1781–1792.

Charleston, M. A. 1998. Jungles: A new solution to the host/parasite phylogeny reconciliation problem. Math. Biosci. **149**:191–223.

Delsuc, F., H. Brinkmann and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. **6**:361–75.

Dress, A. W. M. and D. H. Huson. 2004. Constructing splits graphs. IEEE T. Comp. Biol. Bioinf. **1**:109–115.

Durand, D., B. V. Halldorsson and B. Vernot. 2005. Hybrid micro-macroevolutionary approach to gene tree reconstruction. In RECOMB, vol. 3500 of *Lect. Notes Comput. Sc.*, 250–264.

Efron, B., E. Halloran and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sc. **93**:13429–13434.

Esser, C., N. Ahmadinejad, C. Wiegand, C. Rotte, F. Sebastiani, G. Gelius-Dietrich, K. Henze, E. Kretschmann, E. Richly, D. Leister, D. Bryant, M. A. Steel, P. J. Lockhart, D. Penny and W. Martin. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol. **21**:1643–60.

Felsenstein, J. 1985. Confidence-limits on phylogenies, an approach using the bootstrap. Evolution **39**:783–7911.

———. 2004a. Inferring Phylogenies. Sinauer Associates Inc.

———. 2004b. PHYLIP 3.6: the phylogeny inference package. URL `http://evolution.genetics.washington.edu/phylip.html`.

Fitch, W. M. 1997. Networks and viral evolution. J. Mol. Evol. **44**:S65–S75.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704.

Gusfield, D. and V. Bansal. 2005. A fundamental decomposition theory for phylogenetic networks and incompatible characters. In RECOMB, vol. 3500 of *Lect. Notes Comput. Sc.*, 217–232.

Hallett, M. and J. Lagergren. 2000. New algorithms for the duplication-loss model. In RECOMB, 138–146.

Hedges, S. B. 2002. The origin and evolution of model organisms. Nat. Rev. Genet. **3**:838–849.

Hein, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98**:185–200.

Hendy, M. D. and D. Penny. 1993. Spectral analysis of phylogentic data. J. Classif. **10**:5–24.

Holland, B., K. T. Huber, V. Moulton and P. Lockhart. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol. Biol. Evol. **21**:1459–1461.

Holmes, S. 2005. Statistical approach to tests involving phylogenies. In O. Gascuel, ed., Mathematics of Evolution and Phylogeny, 91–120. Oxford University Press.

Huber, K. T., M. Langton, D. Penny, V. Moulton and M. Hendy. 2002. Spectronet: A package for computing spectra and median networks. Appl. Bioinf. **1**:2041–2059.

Hudson, R. R. 1983. Properties of the neutral allele model with intergenic recombination. Theor. Popul. Biol. **23**:183–201.

Huson, D. H. 1998. SplitsTree: A program for analyzing and visualizing evolutionary data. Bioinformatics **14(10)**:68–73.

Huson, D. H., T. Dezulian, T. K. Kloepper and M. A. Steel. 2004. Phylogenetic super-networks from partial trees. IEEE T. Comp. Biol. Bioinf. **1**:151–158.

Huson, D. H. and T. H. Kloepper. 2005. Computing recombination networks from binary characters. In ECCB, (in press).

Huson, D. H., T. H. Klöpper, P. J. Lockhart and M. A. Steel. 2005. Reconstruction of reticulate networks from gene trees. In RECOMB, vol. 3500 of *Lect. Notes Comput. Sc.*, 233–249.

Jones, D. T., W. R. Taylor and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–82.

Kolaczkowski, B. and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**:980–984.

Legendre, P. and V. Makarenkov. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst. Biol. **51**:199–216.

Linder, C. R. and L. H. Rieseberg. 2004. Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. **91**:1700–1708.

Lyngsoe, R. B., Y. S. Song and J. Hein. 2005. Minimum recombination histories by branch and bound. In WABI.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. **46**:523–536.

Morrison, D. 2005. Networks in phylogenetic analysis: new tools for population biology. Int. J. Parisitol. **35**:567–582.

Nakhleh, L., T. Warnow and C. R. Linder. 2004. Reconstructing reticulate evolution in species - theory and practice. In RECOMB, 337–346.

Philippe, H., N. Lartillot and H. Brinkman. 2005. Multigene analysis of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. **22**:1246–1253.

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

Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. **43**:304–311.

Ronquist, F. and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572–4.

Saitou, N. and M. Nei. 1987. The Neighbour-Joining method: a new method for reconstruction of phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Scheffé, H. 1953. A method for judging all contrasts in the analysis of variance. Biometrika **40**:87–104.

Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang and J. I. Mullins. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol. **73**:10489–502.

Sokal, R. R. and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Science Bull. **38**:1409–1438.

Spencer, M., E. Susko and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. **22**:1161–4.

Steel, M. A. 2005. Should phylogenetic models be trying to ‘fit an elephant’? Trends Genet **21**:307–309.

Steel, M. A., M. D. Hendy, L. A. Székely and P. L. Erdős. 1992. Spectral analysis and a closest tree method for genetic sequences. Appl. Math. Lett. **5**:63–67.

Swofford, D., G. J. Olsen, P. J. Waddell and D. M. Hillis. 1996. Phylogenetic inference. In D. M. Hillis, C. Moritz and B. K. Mable, eds., Molecular Systematics, 407–514. Sinauer, 2nd edn.

Templeton, A. R., K. A. Crandall and C. F. Sing. 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.

Tukey, J. 1953. The problem of multiple comparisons. Unpublished preprint.

Wetzel, R. 1995. Zur Visualisierung abstrakter Ähnlichkeitsbeziehungen. Ph.D. thesis, Fakultät Mathematik, Universität Bielefeld.

Winkworth, R., D. Bryant, P. J. Lockhart, D. Havell and V. Moulton. 2005. Biogeographic interpretation of splits graphs: least squares optimization of branch lengths. Syst. Biol. **54**:56–65.

Wolf, Y. I., I. B. Rogozin and E. V. Koonin. 2004. Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. **14**:29–36.

Appendix

Bootstrap approximate confidence networks

A confidence network is essentially a representation of multiple confidence intervals: one confidence interval for the weight of each split. The problem of constructing confidence networks is therefore the same as the problem of constructing *simultaneous confidence intervals*. Methods for this problem date back to (Scheffé, 1953; Tukey, 1953).

Suppose that for each split we construct the interval for the weight of the split. We say that the intervals are *balanced* if the probability that each interval contains the true weight is the same for each interval. We say that the collection of intervals (or equivalently, the confidence network) *level* $1 - \alpha$, if the probability that *all* of the intervals contain the true weights simultaneously is $1-\alpha$. The goal when constructing simultaneous confidence intervals is to construct a balanced collection of intervals with the correct level.

In SplitsTree4 we have implemented the non-parametric bootrapping 'B method' of (Beran, 1988, 1990) to construct the simultaneous confidence intervals represented by a confidence network; see (Beran, 1988) for a complete description of the method (Note that we use the difference between split weights as the confidence set root when implementing the B method). Beran proved that, under fairly general conditions, the confidence sets will be asymptotically correct to a first order approximation. Hence, given sufficient bootstrap replications and long enough sequences, the confidence network will have close to the correct level and will be approximately balanced.

There are two major caveats. Firstly, network methods like neighbor-net are not continuous, in the sense that a small change in the data can sometimes cause a substantial change in the inferred network. This means that the asymptotic convergence conditions outlined in (Beran, 1988) will hold locally, but not over the entire parameter space.

The second problem is that methods like split decomposition and neighbor-net only construct split networks involving a small number of splits, compared to the number of splits in total. When bootstrapping, many splits appear in only one or two bootstrap replicate networks and so most splits have weight zero for almost all replicates. The net result is that, in simulation, the confidence networks constructed using Beran's B method have incorrect level, even with sequences of length 5000 and 1000 bootstrap replicates. The problem is caused by splits that appear in the true tree but not in the estimate tree, perhaps because of their signals being lost in the sampling error.

There are several avenues for future investigation. Beran (1990) describes a *double* bootstrap method that is more accurate than the original B method. (Efron et al., 1996) applied a double bootstrapping to one dimensional hypothesis tests in phylogenetics. Unfortunately, both of these methods require that the number of bootstrap replicates be squared.

Network tree-likeness test

Suppose that we have inferred a split network $\hat{\mathcal{N}}$ from some data and that this network is not a tree. We want to test whether it is likely that the data originated from a tree. This is easy to do if we have an efficient and correct method for constructing correct confidence networks.

(a) Construct a confidence network for $\hat{\mathcal{N}}$ with level $1 - \alpha$.

(b) If the confidence network does not "contain" a tree (in the technical sense defined above) then reject the null hypothesis that the data originated on a tree.

Part (b) can be executed quickly by constructing the set of splits with confidence intervals excluding zero and rejecting the null hypothesis if and only if this set is incompatible. The performance of the test will, naturally, depend on the confidence network method used. Simulations using the above confidence network indicate that the test is correct but has unacceptably low power. Much work remains.

The architecture of SplitsTree4

The graphical interface makes it easy for a user to interactively explore their data using different phylogenetic methods. The user can select from different analyses, filter data, and manipulate networks, all using standard menus. The data, and networks, can be imported and exported using a variety of different file formats. Multiple analyses can be run simultaneously and the program is multi-threaded so time-consuming calculations are done in the background. SplitsTree4 uses the NEXUS format, as does PAUP*, Mr Bayes, Mesquite, and MacClade.

For the more advanced features, it is important to have an understanding of how SplitsTree4 organizes and processes data. The program arranges its data into a list of different *blocks*, namely the *taxa*, *unaligned*, *characters*, *distances*, *quartets*, *trees*, *splits* and *network* blocks. The taxa block is mandatory, the other blocks are optional. The list of blocks, in this order, is called the *processing pipeline*. Additional blocks are used to direct the analyses and assumptions made by the program.

Usually, the initial input to the program will consist of a taxa block and one additional block containing the input data, which is called the *source* block. Any phylogenetic method is viewed as a *transformation* of one type of data block to another. Most analyses involve a sequence of transformations applied in the same order as the processing pipeline. For example, suppose that the source is a characters block containing an alignment of DNA sequences and the goal is to produce the neighbor-net network for this data. The data in the characters block is first transformed into a distances block using a distance calculation such as the *uncorrected-P* calculation. Next, the distances block is transformed into a splits block using the neighbor-net algorithm. Finally, the data in the splits block is transformed into a network block using a network construction algorithm such as "equal angle" (Dress and Huson, 2004).

SplitsTree4 is written in Java and installers are available for Windows, Mac OS X, and Unix/Linux. The program can be used interactively in a GUI mode, or can be run in a non-interactive mode using the command line to facilitate batch processing.

Figures

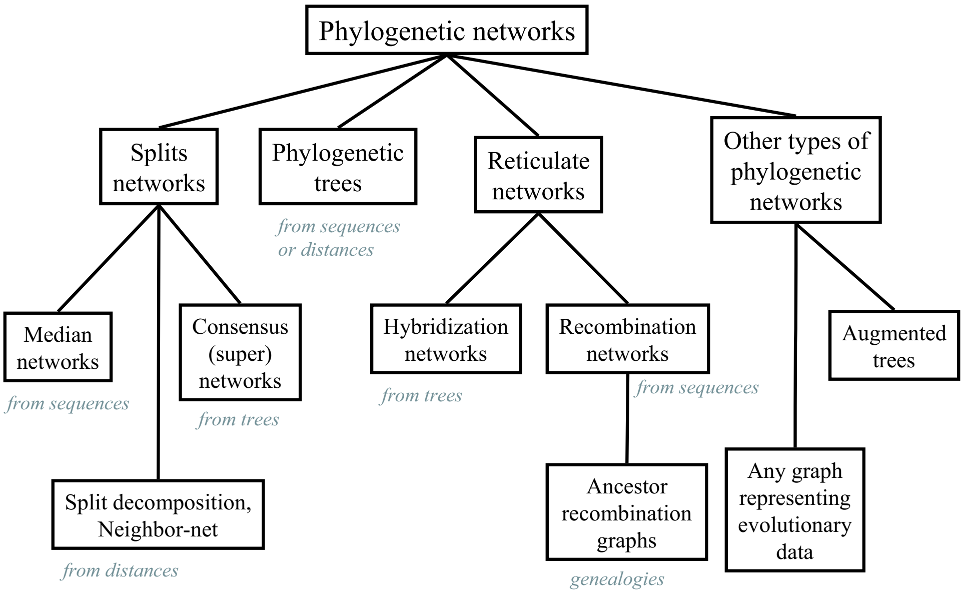

Figure 1:

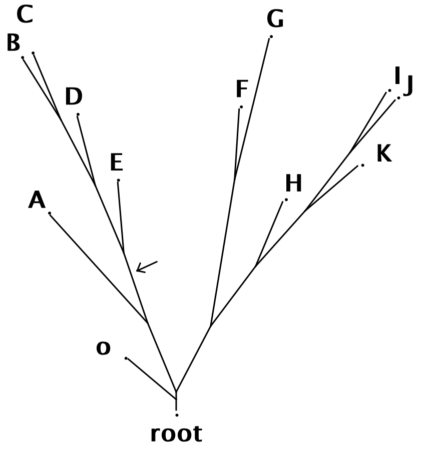

Figure 2(a):

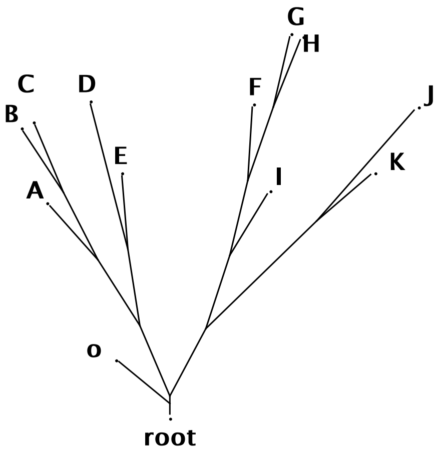

Figure 2(b):

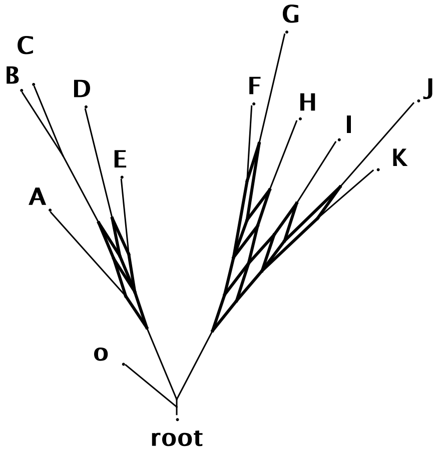

Figure 3(a):

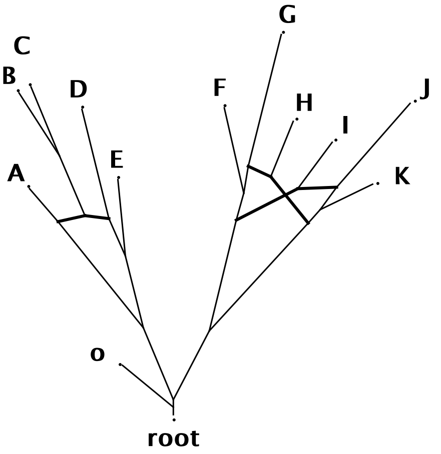

Figure 3(b):

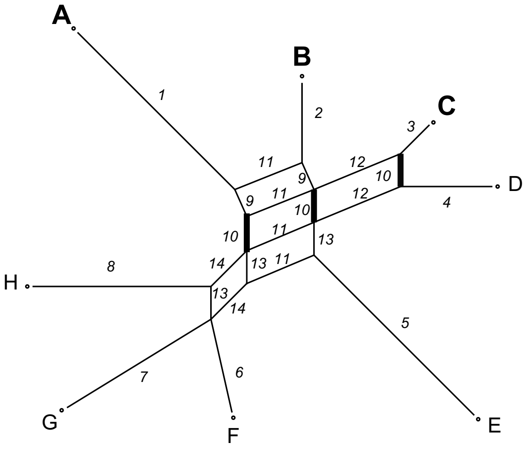

Figure 4(a):

Figure 4(b):

	A	B	C	D	E	F	G	H	weights
1	●	○	○	○	○	○	○	○	7.92
2	○	●	○	○	○	○	○	○	3.31
3	○	○	●	○	○	○	○	○	1.74
4	○	○	○	●	○	○	○	○	3.72
5	○	○	○	○	●	○	○	○	8.94
6	○	○	○	○	○	●	○	○	3.88
7	○	○	○	○	○	○	●	○	5.63
8	○	○	○	○	○	○	○	●	6.21
9	●	●	○	○	○	○	○	○	1.12
10	●	●	●	○	○	○	○	○	1.28
11	○	●	●	●	●	○	○	○	2.83
12	○		●	●	○	○	○	○	3.63
13	○	○	○	○	●	●	●	○	1.28
14	○	○	○	○	○	●	●	●	1.95

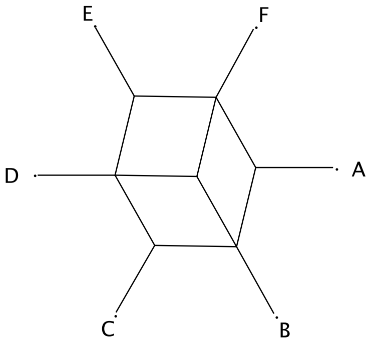

Figure 5(a):

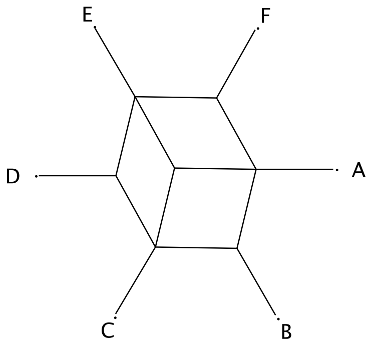

Figure 5(b):

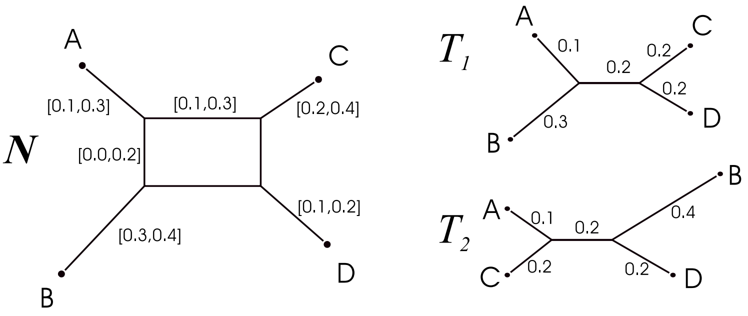

Figure 6:

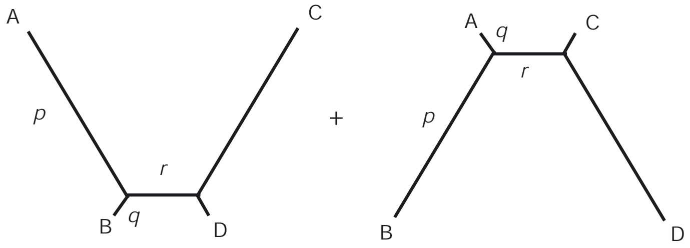

Figure 7(a):

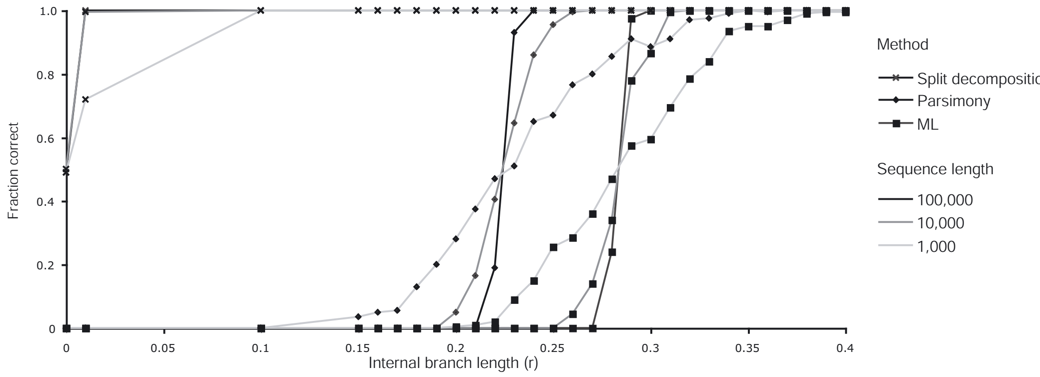

Figure 7(b):

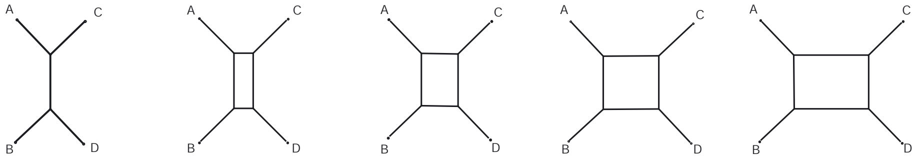

Figure 7(c):

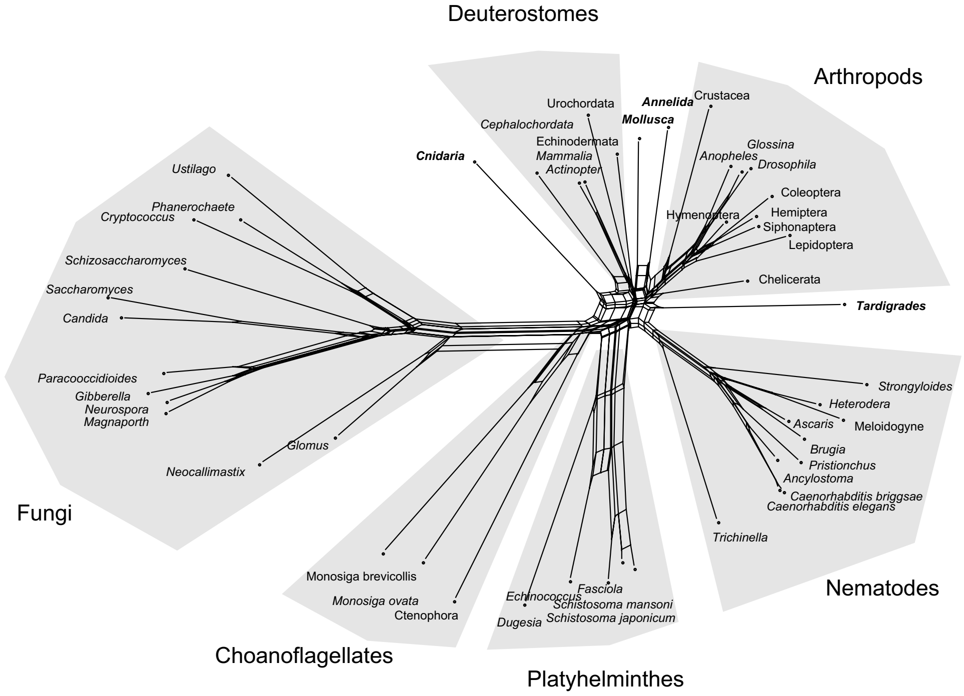

Figure 8:

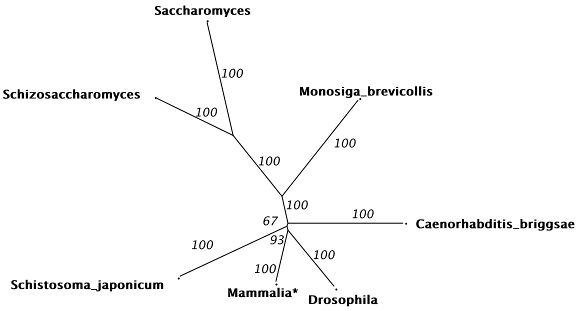

Figure 9(a):

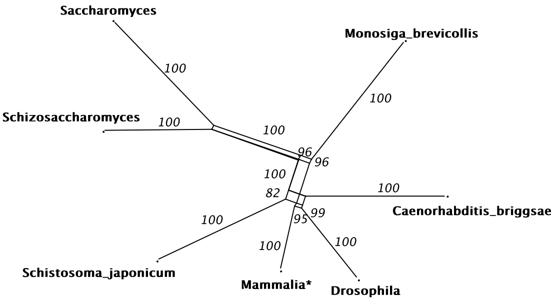

Figure 9(b):

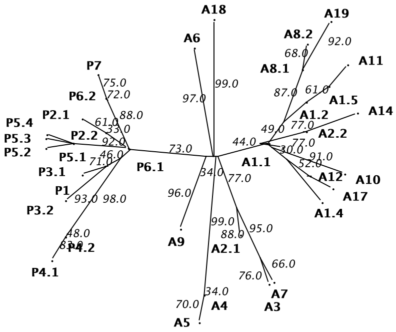

Figure 10(a):

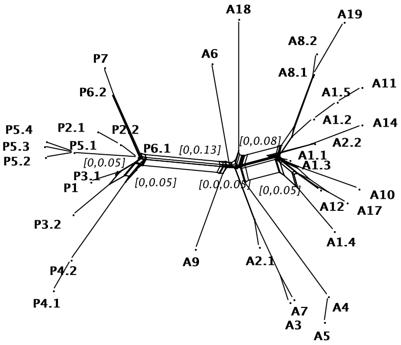

Figure 10(b):

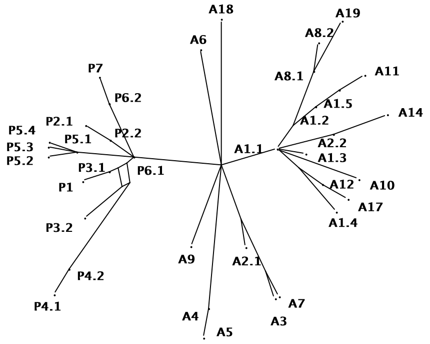

Figure 10(c):

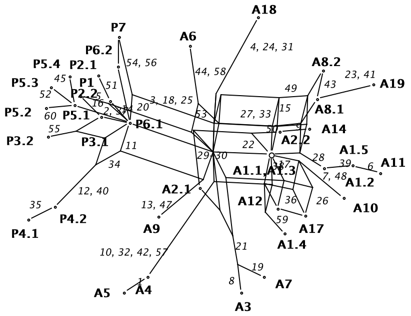

Figure 10(d):

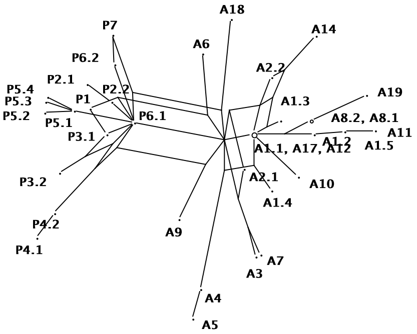

Figure 10(e):

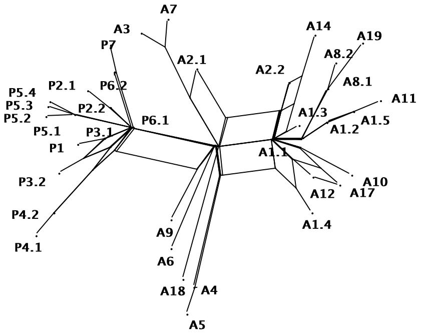

Figure 10(f):

Captions

Figure 1: The term *phylogenetic network* encompasses a number of different concepts, including *phylogenetic* trees, *split networks*, *reticulate networks*, the latter covering both *hybridization* and *recombination* networks, and other types of networks such as *augmented trees*. Recombination networks are closely related to *ancestor recombination graphs* used in population studies. Split networks can be obtained from character sequences e.g. as a median network, from distances using the split-decomposition or neighbor-net method or from trees as a consensus network or super network. Augmented trees are obtained from phylogenetic trees by inserting additional edges to represent e.g. horizontal gene transfer. Other types of phylogenetic networks include host-parasite phylogenies or haplotype networks. Diagram adapted from (Huson and Kloepper, 2005)

Figure 2: Two different trees on the taxon set $X = \{o, A, B, \ldots, K\}$.

Figure 3: Two different types of phylogenetic networks. (a) A split network representing all splits present in the two trees depicted in the previous figure. Here, each band of parallel edges corresponds to a branch contained in one of the input trees. The nodes do not necessarily correspond to hypothetical ancestors. (b) A reticulate network that explains the two trees by postulating three reticulations that give rise to the clades $\{B, C\}$, $\{H\}$ and $\{I\}$. This network explicitly describes a putative evolutionary history: the internal nodes correspond to ancestral taxa and the edges represent patterns of descent.

Figure 4: Two representations of the same information. (a) A split network representing the diversity of HIV strains in a single patient at a single timepoint, data from (Shankarappa et al., 1999). For clarity, edges are labeled by the number of the associated splits. Additionally, all edges representing split number 10, and all taxa on one side of this split, are highlighted in bold. (b) The same information presented as a table of splits and weights (for example, representing the average number of substitutions per site) given for each split. Each split divides the taxa into two groups, one represented by open circles, the other by closed circles. For example, split 10 divides the taxa into $\{A, B, C\}$ and $\{D, E, F, G, H\}$.

Figure 5: Two different representations of the same set of splits.

Figure 6: A simple confidence network N and two trees T_1, T_2. The network contains the tree T_1 with the given branch lengths, but it does not contain tree T_2 since the split $AB|CD$ has weight 0 in tree T_2 (that is, it is not present) but the confidence interval for the weight of $AB|CD$ in the network does not include 0.

Figure 7: The effect of rate heterogeneity on ML, MP, and split decomposition. (a) Following the model proposed in (Kolaczkowski and Thornton, 2004), sequences are evolved on a four taxa tree, half with one set of branch lengths and half with the other. Here, $p = 0.75$, $q = 0.05$ and r varies between 0 and 0.4 expected mutations per site. Sequences were simulated, and analyzed, using a Jukes-Cantor model. (b) For each internal branch length r, the fraction of 200 replicates that each method returned the correct tree or a network containing the correct tree. (c) The split networks corresponding to infinite sequences for internal branch values $r = 0, 0.1, 0.2, 0.3, 0.4$.

Figure 8: A neighbor-net constructed from a concatenation of 71 genes from 49 animals, fungi and choanoflagellates. The major groupings are indicated. The network does not conclusively support either the coelomate hypothesis (molluscs, deuterosomes, arthropods grouping together) or the ecdysozoa hypothesis (arthropods, nemotodes, tardigrades grouping together) but suggests that there is evidence for both. A network tree-likeness test indicates that the conflict is not merely the product of sampling error. Instead the network is representing the effect of problems with the model or biases in the estimation methods.

Figure 9: (a) The Bio-NJ tree (Gascuel, 1997), with bootstrap values, for a smaller set of 7 animals (following (Philippe et al., 2005)) using a concatenated alignment of 146 genes, ML distances under a JTT+F+Γ model. The tree-based method gives reasonable, but not conclusive, support for the coelomate hypothesis, though the small bootstrap value, even with this large number of sites, already suggests that the clade is unreliable. (b) The neighbor-net network using the same distance estimates, with bootstrap values. Even without the cnidarian taxa, there is substantial (but not conclusive) support in the data for the ecdysozoa hypothesis.

Figure 10: For an alignment of 60 variables sites of DNA for 35 haplotypes of dusky dolphins, we show: (a) The neighbor-joining tree with bootstrap values. (b) The 95% confidence network obtained from the preceeding tree. To avoid clutter, we only show confidence intervals for the main internal splits. (c) The consensus network of the three most parsimonious trees. (d) The median network, with edges labeled by supporting sites. (e) The split decomposition network. (f) The neighbor-net network.