Available online at www.sciencedirect.com

SCIENCE DIRECT®

International Journal for Parasitology 35 (2005) 567–582

www.parasitology-online.com

Invited review

Networks in phylogenetic analysis: new tools for population biology

David A. Morrison*

Department of Parasitology (SWEPAR), National Veterinary Institute and Swedish University of Agricultural Sciences, 751 89 Uppsala, Sweden

Received 31 January 2005; received in revised form 10 February 2005; accepted 10 February 2005

Abstract

Phylogenetic analysis has changed greatly in the past decade, including the more widespread appreciation of the idea that evolutionary histories are not always tree-like, and may, thus, be best represented as reticulated networks rather than as strictly dichotomous trees. Reconstructing such histories in the absence of a bifurcating speciation process is even more difficult than the usual procedure, and a range of alternative strategies have been developed. There seem to be two basic uses for a network model of evolution: the display of real but unobservable evolutionary events (i.e. a hypothesis of the true phylogenetic history), and the display of character conflict within the data itself (i.e. a summary of the data). These two general approaches are briefly reviewed here, and the strengths and weaknesses of the different implementations are compared and contrasted. Each network methodology seems to have limitations in terms of how it responds to increasing complexity (e.g. conflict) in the data, and therefore each is likely to be more appropriate for one of the two uses than for the other. Several examples using parasitological data sets illustrate the uses of networks within the context of population biology.

Keywords: Phylogeny; Phylogenetic networks; Reticulograms; Statistical parsimony; Median networks; Split decomposition; Neighbour-net; *Dictyocaulus*; *Sarcoptes*; *Toxoplasma*

1. Introduction

Biologists are confronted with complexity on a scale undreamt-of by physicists and chemists, so much so that biologists have devised the unique word ‘biodiversity’ to describe it. They have traditionally dealt with the study of this complexity by first extracting from it the simplest pattern that they can find, and then gradually adding more and more complexity to this simple model in order to increase the predictive and/or explanatory power. At no stage do they necessarily expect to arrive at a complete description of the true complexity, in the same sense that physicists and chemists seem to have traditionally expected. Charles Darwin (1859) first pointed out that the simplest pattern that might lie at the heart of evolutionary history can be represented by a tree; indeed, the only illustration in that book uses a tree to present the idea of phylogenetic relationships. Biologists have, therefore, been very interested in mathematical algorithms for deriving trees from character data, in the expectation that the trees might represent a good starting point for studies of evolution. These trees are usually interpreted in biological terms, as representing ancestors (inferred) and descendants (possibly observed) along with character-state changes between them, so that the putative evolutionary history is displayed. The next step is then to add some complexity to this simplistic model.

To this end, mathematicians and mathematically inclined biologists have, over the past 15 years (although initial interest long pre-dates that time; Sneath, 1975), been developing algorithms for networks, which are the most obvious generalisations of a tree model for evolution. These networks allow reticulations among the branches on the diagram, rather than imposing a strictly bifurcating structure—compare Fig. 1a and b for an example. A mathematical network (i.e. a tree with reticulations) can have two potential interpretations in a biological context, and the distinction between them has not been sufficiently emphasised in the literature: (i) a representation of character-state evolution amongst ancestors and descendants that have complex relationships due to non-dichotomous historical events (i.e. an hypothesis of the true

* Tel.: +46 18 67 4161; fax: +46 18 30 9162.
E-mail address: david.morrison@bvf.slu.se.

0020-7519/$30.00 © 2005 Australian Society for Parasitology Inc. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijpara.2005.02.007

phylogenetic history); and (ii) a representation of character conflict or uncertainty within the dataset, where some characters show non-compatible patterns (i.e. simply a data summary). Strictly speaking, only the first sort of network should be called a ‘phylogenetic network’, although many people have used this term for the second sort as well.

There are a number of reasons why biologists might believe that true evolutionary relationships are reticulate rather than strictly tree-like (i.e. interpretation (i) above). For example, at the population level, biologists have always thought of relationships among individuals as being reticulate, simply because the primary historical pattern within a sexually reproducing species is inter-breeding. This clearly creates an anastomosing network of relationships among parents and offspring. That is, a ‘family tree’ is not actually a tree at all, not even in the colloquial sense, as inter-marriages create inter-connections among the family members that cannot be represented by a simple hierarchical diagram. Nevertheless, the history of a single gene copy at a single locus (literally, a genealogy) will be a tree, in the absence of confounding population processes. Unfortunately, there are plenty of population-level phenomena that create reticulate genealogical relationships, such as recombination, gene conversion, reassortment (lineage sorting), deep coalescence, etc. (Posada and Crandall, 2001). Thus, the proper model for representing the evolution of molecular sequences is a family of trees, each one describing the history of a segment of the sequence (Hein, 1990). This may explain why it is only recently that biologists have begun to study population-level histories other than for domesticated organisms (where there may be written genealogical records), because reconstructing a population history can be a daunting task.

Above the species level, historically there have also been many depictions of relationships as reticulating, particularly in botany (Stevens, 1994) and microbiology (Sneath, 2000), and more recently for the origin of major taxonomic groups (Doolittle, 1999). Such phylogenetic relationships may truly be reticulate as a result of: hybridisation (e.g. common in plants, but also well known among invertebrates, fish and amphibians); horizontal gene transfer (e.g. very common among bacteria); and symbiosis (e.g. the origin of eukaryote organelles). Furthermore, there will often be multiple phylogenetic trees for the same taxa, which may be discordant, as occurs for example when there is an incompatibility between some or all of the gene trees and the species tree (Smouse, 2000; Posada and Crandall, 2001).

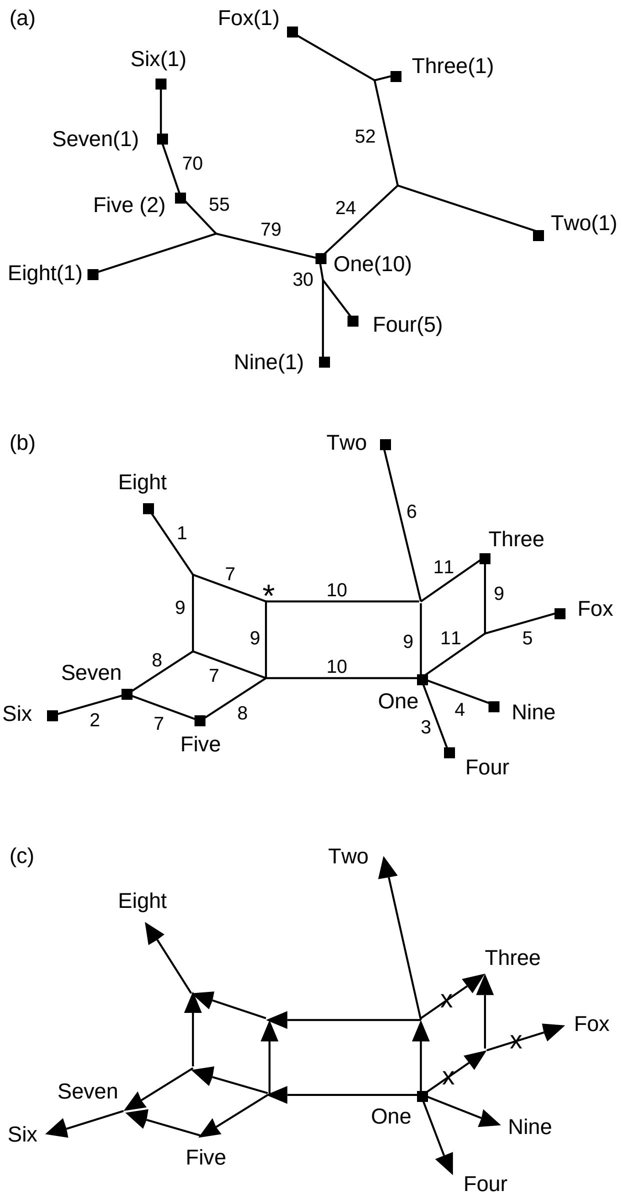

Fig. 1. Phylogenetic analyses of haplotype data for *Sarcoptes scabiei* from wombats, dogs and humans in Australia and a fox in Sweden, based on mitochondrial small-subunit rRNA sequences. (a) Inferred phylogenetic tree from a neighbour-joining analysis using the HKY substitution model. Each of the 10 haplotypes is marked as a box at the appropriate node, and the number of sequences for each haplotype is indicated in brackets. The branch lengths represent the amount of character-state change occurring on that branch. If a root is added to the tree then each branch would have an evolutionary direction away from the root and each unmarked node would represent an inferred (unobserved) ancestor. The numbers on the branches are bootstrap percentages based on 1000 replicates. (b) Inferred character-display network for the same data. The network is built up of parallelograms forming bands of parallel edges, which are numbered 1–11. Each set of parallel edges represents a split (or bipartition) of the 10 named haplotypes into two non-overlapping groups, with each edge length being proportional to the number of nucleotide characters that support that particular split. Thus, each numbered edge represents a particular pattern of character states. The tree-like parts of the diagram represent uncontradicted patterns of character change, while conflicting character patterns create box-like structures. In order to turn this network into a dichotomous tree, at least one edge from each of the four boxes needs to be ‘cut’. The various possible combinations of these cuts represent the possible dichotomous trees that are compatible with the data. The asterisk marks the node (and three incident branches) that do not appear in the network from the statistical parsimony analysis. (c) Directed acyclic graph resulting when the network shown in (b) is rooted using haplotype One as the ancestor. Each branch has an evolutionary direction away from the root, as indicated by the arrows. The crosses mark the three branches whose direction would reverse if the Fox haplotype is chosen as the root instead.

There may also be situations where trees from different sources are being compared, such as in the study of biogeography or host-parasite relationships, and a reconciliation of the various trees is required (Legendre, 2000b).

Here, I focus my attention and examples mainly on the use of networks in population-level studies, as being more appropriate for an introduction to the topic. Such population-level analyses can differ in several ways from species-level analyses (Posada and Crandall, 2001), most notably in that many of the internal nodes on a phylogenetic diagram (tree or network) may be contemporary, so that ancestors appear in the data set with their descendants. (Technically, this makes the diagram a spanning tree rather than a steiner tree.) However, they also often involve low levels of divergence among the taxa (perhaps with many identical taxa) and large sample sizes.

A character display that directly incorporates conflicts among the character-state patterns into the visual representation (i.e. interpretation (ii) above) is clearly a valuable analytical tool. Phylogenetic trees can be unstable because the conflicting character patterns are not consistent with a single tree, but instead support several mutually incompatible trees. This is why different tree-building methods often produce different trees for the same data set and why some methods produce many equally optimal trees. Currently, the most common way of indicating this sort of character conflict in an evolutionary history is to put numbers on the branches of a single preferred tree (see Fig. 1a). These are intended to represent the amount of 'support' for that branch (e.g. bootstrap proportions), so that small numbers represent lack of support, which is usually then inferred to be caused by conflicting patterns among the characters. However, this approach provides no simple visual clue as to the location of conflicts (i.e. all numbers visually look pretty much the same). More importantly, it does not in any way show what the alternative (i.e. conflicting) patterns look like, because only the predominant pattern appears on the tree Fig. 1a. A network is designed to fill both roles, since both the location and the nature of character conflicts are clearly indicated by inter-connections among the branches (see Fig. 1b).

So, in an ideal network display of character conflict, tree-like areas of the diagram would represent those parts of the phylogeny with no conflict among the characters, areas with multifurcations would represent those parts of the phylogeny where there is not sufficient data to reconstruct the phylogeny, and those areas with reticulations would represent those parts of the phylogeny where two or more of the character-state patterns conflict with each other. Therefore, the more tree-like the data are then the more tree-like will be the network, which gives an instant visual clue to how many alternative bifurcating trees are compatible with the data.

As is usual for phylogenies, the lengths of the branches should represent the number of character-state changes or the genetic distance, so that the relative length of the anastomoses indicates the degree to which the alternative patterns are supported by the data. The sum of the lengths of the branches along the shortest path connecting two taxa should be proportional to the actual distance between those taxa, although for a network there will be multiple shortest paths (unlike a tree where there is only one path connecting any pair of taxa). For example, in Fig. 1b there are three equally short paths from haplotype Two to haplotype Seven, using branches $6\rightarrow10\rightarrow7\rightarrow9\rightarrow8$ or $6\rightarrow9\rightarrow10\rightarrow8\rightarrow7$ or $6\rightarrow9\rightarrow10\rightarrow7\rightarrow8$, whereas there would only be one such path in a tree.

2. Different ways to construct networks

Two rather different approaches have been taken to developing mathematical networks to fill the two roles described above (i.e. reconstructing phylogenetic history or displaying character conflict). This is quite different to the situation regarding phylogenetic trees. The usual way to construct a tree is first to display the character-state data and then to turn the resulting data-summary diagram into an evolutionary diagram by adding a root. That is, if there is no conflict among the characters then the characters and their states will form a perfect nested hierarchy, and this can best be displayed as a tree. Mathematically, this is an undirected connected line graph. If we then root the tree (using external information, e.g. by specifying that one of the taxa is an outgroup) this gives an evolutionary direction to the tree, so that each branch and node on the tree is part of a unique and unambiguous path away from the root. Mathematically, this now a directed acyclic graph (where a cycle is a directed path from a node back to itself, which would be nonsense in a phylogenetic context because it would imply that a descendant could be its own ancestor). We can then interpret this (rooted) directed tree in terms of character-state evolution among ancestors and descendants (i.e. a phylogenetic tree where each internal node is an ancestor and each leaf is a contemporary descendant). So, character display and phylogenetic interpretation are inter-twined in a tree diagram. This is possible because the tree-building method effectively ignores the character conflict (or, more accurately, treats it as uninteresting ambiguity), and thus tries to display only the 'true' evolutionary events.

Networks can potentially be constructed in the same way; that is, by first displaying the patterns and then trying to turn the display into an evolutionary diagram. In this case, conflict among the characters will produce a data summary that forms an anastomosing plexus instead of a tree. Mathematically, this is an undirected cyclic graph. If we root this network (e.g. by specifying an outgroup) then this gives an evolutionary direction to the network, and each branch on the network will also have a direction (although this may sometimes be ambiguous) (see Fig. 1c). This is now a directed acyclic graph (i.e. the directions on the branches associated with reticulations should not lead back to themselves). This is the approach taken by all of the 'character-display' forms of network, which are, therefore, effective (to varying degrees) as character displays.

Unfortunately, a directed network formed in this way cannot be interpreted in terms of character-state evolution among ancestors and descendants. This is because not all of the nodes in the network necessarily represent ancestors nor do all of the reticulations necessarily represent evolutionary events (Nakhleh et al., 2003; Bryant and Moulton, 2004; Winkworth et al., in press). The nodes and branches represent character conflict, and not every piece of character conflict in the data necessarily represents a real evolutionary event. Thus, some of the nodes and branches may be 'false positives' if they are interpreted as representing real phylogenetic history. Such a network should not be interpreted as a true phylogenetic network, at least not without some extra work. This ought not really be a problem, because these types of networks were only ever designed as character displays and not as displays of the hypothesised true phylogeny—that is, they are best used as exploratory tools rather than as definitive phylogenetic analyses. In this sense they have much in common with consensus trees, which are mathematical summaries of multiple phylogenetic trees without any necessary role in depicting the true phylogeny (Bryant, 2003).

To meet this limitation, alternative methods for constructing explicit phylogenetic networks have been developed. These methods try to model the various biological processes that can form reticulations in phylogenetic trees (e.g. recombination, hybridisation, lateral gene transfer) by detecting the various character patterns that these different biological processes can create in the character data. That is, a basic rooted-tree model is assumed, to which directed reticulations are added, thus evolving ancestral sequences into descendant sequences according to some mathematical model of the particular biological process that is being assumed for the data. Temporal ordering is explicitly incorporated into the construction of these networks, so that they are rooted directed acyclic graphs by construction, with the root (i.e. the ancestral taxon) known a priori. These are, thus, true phylogenetic networks in the sense that the method explicitly tries to construct a diagram that can be interpreted as a phylogeny, rather than simply being a character display.

2.1. True phylogenetic networks

This methodology has so far only been applied to molecular data, rather than to other organismal features. The underlying idea is that some basic unit of the data will show tree-like relationships even though the data as a whole are properly represented by a network (Hein, 1990; Linder and Rieseberg, 2004). For example, any one segment of linked nucleotides in a gene will have evolved down a single evolutionary tree, even though recombination or hybridisation has caused different segments of the gene to have different evolutionary trees. So, the basic phylogenetic model for every nucleotide is a tree, and each organism is represented by a family of phylogenies. The objective of the network analysis is to combine the various trees into the optimal network representing that collection of trees, thus allowing us to infer the origin of the recombination, hybridisation, or lateral transfer events. That is, the minimum number of reticulation events is inferred to be the most likely (i.e. most parsimonious) history (Hein, 1990).

Several different approaches to producing this type of explicit phylogenetic network have been developed. One way to think about the relationship between networks and trees is to recognise that each recombination, hybridisation or lateral gene transfer event will create two phylogenetic trees that differ from each other in only one sub-tree (Hein, 1990). That is, one branch of the tree, along with all of the branches descending from that branch, is moved from its original location to a new location (the branch movement represents the transfer of genetic material). Mathematically, the trees differ by one rooted subtree-prune-regraft operation. This means that a network can be constructed by minimising the number of these operations needed to reconcile the observed trees, and several algorithms have been proposed to do this (Hein, 1990, 1993; Nakhleh et al., 2004).

However, there are serious combinatorial problems involved in implementing this type of analysis. That is, searching for an optimal solution is even more complex in network analysis than it is in tree-building, which is well known for being mathematically intractable in many cases. All of the models are, thus, rather restricted in one way or another, making what may be biologically unrealistic simplifying assumption in order to make the mathematics tractable. In particular, they mostly assume that the reticulation events are relatively rare, that the data sets are small (i.e. only a few gene trees), or that the data consist of binary characters (which may not be much help if you have DNA sequence data).

Most of the recent methods are based on the idea of using a directed acyclic graph as a model for describing an evolutionary history with reticulations. Different properties or conditions are specified for the graph model in order for it to provide a realistic biological model as well. For example, conditions have been described to allow the graphs to represent recombination (Strimmer and Moulton, 2000), lateral gene transfer (Hallett and Lagergren, 2001; Hallett et al., 2004), or hybridisation (Moret et al., 2004; Nakhleh et al., 2004), which might leave different 'signatures' in the data and therefore produce different patterns in the graph. Since the graph is directed (i.e. there is a root and all of the branches have an unambiguous direction), three distinct types of nodes can be recognised in the network: tree nodes (one branch coming in and two or more going out), which represent genetic divergence; reticulation nodes (two branches coming in and one going out), representing recombination, hybridisation, lateral transfer, etc.; and the root node (no branches coming in and two or more going out). Thus, all of the nodes represent hypothesised ancestors and the branches represent evolutionary changes (e.g. mutations), so that the network can be interpreted as an attempt to reconstruct the true phylogeny. Because of the time direction of the branches, reticulation events can also act as time synchronisation points, since reticulations can only occur between taxa that are contemporaneous (e.g. recombination cannot occur between a descendant and its ancestor). This places realistic constraints on the form of the network, and can allow events to be dated.

Some of the current implementations of this approach are based on the idea of reconciling gene trees in a species tree (Hallett and Lagergren, 2001; Hallett et al., 2004; Nakhleh et al., 2004). That is, the gene trees should ultimately reflect the true species tree but do not do so as a result of the hypothesised reticulation events. So, a network is constructed that optimally reconciles the conflicts, which will reveal the reticulated history of the species. However, these implementations are still rather limited. For example, only a pair of gene trees can be analysed, and an unlimited number of reticulation events can only be inferred if the species tree is already known, with much more severe restrictions on the number of events if the tree is unknown (i.e. reticulations are assumed to be rare). Another simplification that has been employed is to assume that reticulation loops in the network do not share nodes (i.e. all of the loops are isolated from each other; Wang et al., 2001). This constrained form of network is called a 'galled tree', and it greatly simplifies the calculations and search for an optimal network (Gusfield et al., 2004). However, the only currently released computer program implementing any of these ideas is that of Addario-Berry et al. (2003) (http://cgm.cs.mcgill.ca/laddar/lattrans/), although a program for producing galled trees is also available (http://wwwcsif.cs.ucdavis.edu/gusfield/).

Alternatively, an ancestral recombination graph might be a more appropriate model on which to base a directed acyclic graph (Strimmer et al., 2001). These are rooted graphs that summarise linked collections of ultrametric (i.e. clock-like) trees in a network. The link between the trees, in this context, is that they are a set of dichotomous subtrees derivable from the network—the network is created by 'gluing together' the trees. The advantage here is that this allows a likelihood framework to be applied to network construction. This requires an explicit specification of a model of sequence evolution leading to the reticulation events, but extends the above methods by allowing particular hypotheses to be tested. While this is an appealing approach, the likelihood calculations probably make the method prohibitive (Strimmer and Moulton, 2000; Strimmer et al., 2001).

Another model-based approach is that of Xu (2000), who provides several models for hybridisation that can be employed to construct a network using a least-squares criterion. The method is currently restricted to the use of genetic distances based on gene frequency data, and all possible networks need to be evaluated in order to find the optimal one. So, the method is not yet widely applicable.

2.2. Character-display networks

When interpreting these types of diagrams it is important to remember the point emphasised above, that not all character conflict is necessarily relevant to reconstructing reticulate evolutionary history, and thus the 'display' methods will be prone to false positives if all nodes and branches are interpreted as real evolutionary events, even for relatively simple situations. There are actually three possible interpretations of reticulations in a network: (a) they represent uncertainty or ambiguity; e.g. inaccuracies in the data (sampling error) such as mistaken homology in choosing orthologues and misalignment, or the data do not fit the model used (model heterogeneity) such as failure to account for variation in substitution rates between sites; (b) they represent analogy events in evolution, usually referred to as homoplasy; e.g. convergences (similar characters), parallelisms (multiple origins of characters) and reversals (transformations of characters back to a prior state); or (c) they represent homology events in the phylogenetic history, involving actual gene exchange between unrelated organisms; e.g. recombination, hybridisation, lateral gene transfer. It is not necessarily straightforward to distinguish these three possibilities when looking at a network (Bandelt and Dress, 1992), and this will require some thought on the part of the user. Furthermore, branch lengths mathematically correspond to the weight of the split represented by the branch (that is, a branch splits the taxa into two non-overlapping groups, which will be formed by cutting the branch), but the interpretation of the branch lengths may be different for different types of networks. For example, the weights may represent the amount of character change or they may represent the amount of support for the branch.

Currently available network methods in this class include (Linder and Rieseberg, 2004):

(i) *Adding-to-a-tree*: These methods infer a single optimal phylogenetic tree and then add reticulations to that tree in order to optimise some further mathematical criterion. Examples: Reticulograms (see below); Statistical parsimony (see below); Pyramids (Aude et al., 1999); Weak hierarchies (Bandelt and Dress, 1989).

(ii) *Combining trees*: These methods infer a set of optimal phylogenetic trees and then try to simultaneously display all of them. These methods are all based on constructing one-step trees or minimum spanning networks. Examples: Netting (Fitch, 1997); Molecular variance parsimony (Excoffier and Smouse, 1994); Median networks (see below).

(ii) *Computing splits*: These methods directly quantify the data incompatibilities and then try to display these incompatibilities, without ever explicitly inferring a tree.

Examples: Split decomposition (see below); Neighbour-Net (see below); Statistical geometry (Eigen et al., 1988).

Note that most of these methods are analogous to either parsimony-based or distance-based tree-building methods, rather than to maximum-likelihood methods—the computational complexity of implementing this form of network as a maximum-likelihood method is even worse than it is for tree building (von Haeseler and Churchill, 1993). Furthermore, the parsimony-based methods have mostly been designed with population-level data in mind, while the distance-based methods can be applied to any sort of data. The latter work very rapidly, as do most such methods for tree-building as well, which means that they are practical tools even for large data sets.

Most of these different methods will produce roughly the same network for simple character-state patterns, but there may be major differences among them as the character complexity increases. None of these methods have been thoroughly evaluated in terms of their behaviour with different data sets (although see Lapointe, 2000; Posada and Crandall, 2001; Nakhleh et al., 2003), which is different to the situation for phylogenetic trees, where many such studies exist. However, some behavioural generalisations are now known, and these are discussed in Section 3 for five commonly used methods.

3. Commonly used network methods

3.1. Reticulograms

This is a distanced-based method (Legendre and Makarenkov, 2002; Makarenkov and Legendre, 2004). First a phylogenetic tree is inferred based on the chosen distance measure, and then reticulations are added to this tree in order to optimise some additional criterion (measuring goodness-of-fit to the data). This is based on the idea that a tree is the simplest phylogenetic model and therefore we should not deviate from it any more than is absolutely necessary (e.g. the number of reticulation events is likely to be small compared to the number of dichotomous events). However, the method currently includes at least five different tree-building methods and three optimisation criteria, and so it can be considered as a class of methods rather than a single method.

This method can be effective for certain circumstances, since it is based on a tree and therefore the diagram can be interpreted similarly to a tree (e.g. displaying inferred ancestors). However, it is not clear how/when to choose a particular method from the large list of available options, as these have never been quantitatively assessed. The biological interpretation may sometimes be tricky, as it is not always obvious what the reticulations mean (other than that there are patterns in the data that cannot be displayed on the original tree), and the diagram can be sensitive to the choice of initial tree. All of the branches are drawn in the network even if they have zero length (i.e. a leaf must always be on a pendant edge). The network diagram can become uninterpretable with increasing character-state complexity. Even with a specified root, not all branches will necessarily have an unambiguous direction.

A computer program called T-Rex (Makarenkov, 2001) is available from http://www.labunix.uqam.ca/~makarenv/trex.html.

3.2. Statistical parsimony

This is a parsimony-based method (Templeton et al., 1992). The idea is to sequentially connect taxa in the order of their increasing character-state differences, until the ‘parsimony connection limit’ is reached, which is the limit at which parsimony is a reliable phylogenetic inference method. This method is based on the idea that parsimony is an appealing model for reconstructing evolution but it is known to have problems with distant relationships (due to hidden character-state changes). The analysis is usually based on the raw data, but it can also be applied to the hamming distance (i.e. the observed number of character-state differences).

This method is straightforward to apply, because there are no options. The display of character conflict can be interpreted similarly to that of a parsimony tree (i.e. each branch represents a particular character-state change), but it is likely to be incomplete because some possible indirect connections will not be drawn if more-direct connections already exist. Also, the biological interpretation can be similar to that of a parsimony tree, and ancestral nodes are explicitly inferred. Furthermore, coalescent theory in population genetics allows a root to be inferred without an outgroup (Castelloe and Templeton, 1994). Unfortunately, the method responds to increasing complexity by disconnecting the diagram (i.e. several unconnected networks are produced instead of a single diagram), which means that there will be many false negatives. Even with a specified root, not all branches will necessarily have an unambiguous direction.

A computer program called TCS (Clement et al., 2000) is available from http://darwin.uvigo.es/software/tcs.html.

3.3. Median networks

This is a character-based method (Bandelt, 1994; Bandelt et al., 2000), usually applied to binary data. The idea is to simultaneously display all of the character-state differences among the taxa as separate branches in a network (i.e. one branch for each individual pattern). This approach is based on the idea that visually displaying all of the character differences between taxa will be profitable (i.e. all incompatibility should be displayed), as the diagram is guaranteed to include all of the most-parsimonious trees. This is actually a class of methods, as various major alternatives have been proposed, including reduced median networks (Bandelt et al., 1995), median-joining networks (Bandelt et al., 1999), greedily reduced median networks (Bandelt et al., 2000), pruned median networks (Huber et al., 2001a), local buneman graphs (Huber et al., 2001b), and consensus networks (Holland and Moulton, 2003). Some of these, especially pruned networks, can produce disconnected diagrams, and they do not guarantee to include all of the most-parsimonious trees.

The display of character conflict is good for simple cases, but the diagram becomes increasingly complex as the amount of conflict increases, since all conflict is meant to be displayed and this requires a new dimension for each conflicting pattern. So, the diagram can produce undisplayable hypercubes, and beyond a cube the diagram is uninterpretable. A range of alternative methods has been created in order to deal with complex patterns, by leaving out some of the character conflict. However, it is unclear how to choose from this buffet of options (see the list above), as these have never been quantitatively assessed, and they can produce major differences in the resulting network. A simple biological interpretation is possible for straightforward diagrams, but the inferred nodes will not all represent ancestors, which means that there will be false positives if all of the reticulations are interpreted as evolutionary events. With a specified root, all branches should have an unambiguous direction because parallel edges in the diagram are constrained to all point in the same direction.

There are several programs available, including: Network (Fluxus Technology, 2000) http://www.fluxus-engineering.com/sharenet.htm (median networks, median-joining networks); Spectronet (Huber et al., 2002) http://awcmee.massey.ac.nz/spectronet/ (median networks, reduced median networks, pruned median networks); and SplitsTree4 (Huson and Bryant, 2005) http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome.html (consensus networks).

3.4. Split decomposition

The objective is to simultaneously display all of the least-worst quartets among the taxa. That is, there are three possible unrooted trees for a quartet of taxa, and if only one of them has character support then a single overall tree will result, while a network will result if two or more of the quartets have support. This approach is based on the idea of visually displaying as much of the character data as possible in two dimensions while leaving out the rest (i.e. circular splits are subgraphs of hypercubes and can be displayed in a plane whereas *k*-compatible splits, such as are used in a median network, need the hypercubes). It can be based on the raw data (called parsimony splits; Bandelt and Dress, 1993) or more usually on a distance measure (Bandelt and Dress, 1992).

The display of character conflict is good for simple and moderate cases, but the method responds to increasing character-state complexity by producing uninformative multifurcations (i.e. false negatives). Thus, a multifurcation may represent either ‘insufficient information’ or ‘too much conflicting information’, which is not straightforward to interpret. As complexity decreases the method should converge on the median network. A simple biological interpretation is possible for all cases, but the inferred nodes will not all represent ancestors. It is unclear how to choose from the options available (e.g. distance versus parsimony), although the choice seems to have little effect on the outcome. With a specified root, all branches should have an unambiguous direction because parallel edges in the diagram are constrained to all point in the same direction.

There are a couple of programs available, including: SplitsTree (Huson, 1998) http://www-ab.informatik.uni-tuebingen.de/software/splits/welcome.html; and SplitsTree4 (Huson and Bryant, 2005) http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome.html.

3.5. Neighbour-Net

This is a distanced-based method (Bryant and Moulton, 2002, 2004). It tries to generalise the neighbour-joining tree-building method by ‘delaying’ the connections among taxa, in a manner similar to pyramid clustering. This approach is based on the idea of visually displaying the data in two dimensions (i.e. circular splits can be displayed in a plane) but being more informative than split decomposition.

This method was developed as a compromise between the preponderance of apparent false positives in median networks and the false negatives of split decomposition. It is straightforward to apply, because there are currently few options, other than the choice of a distance measure. The display of character conflict is good for simple to relatively complex cases, and it seems to respond well to increasing complexity. This makes it the best network option for analysing the most complex situations. As complexity decreases it should converge on the median network. A simple biological interpretation is possible for all cases, but the inferred nodes will not all represent ancestors. With a specified root, all branches should have an unambiguous direction because parallel edges in the diagram are constrained to all point in the same direction.

A computer program called SplitsTree4 (Huson and Bryant, 2005) is available from http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome.html.

4. Examples

Reticulation events have been shown to have important practical consequences for the results of phylogenetic analyses, including both recombination (Schierup and Hein, 2000) and hybridisation (McDade, 1992). Network analyses should, therefore, always be considered when there is the possibility that the phylogenetic history being examined has been reticulate rather than dichotomous, even if they are used only as an exploratory tool prior to some more-definitive phylogenetic analysis. However, to date there have been almost no applications of network methods within parasitology (e.g. Morrison et al., 2003). To rectify this situation, here I illustrate a few of the potential uses of networks using some parasitological data sets.

Four examples have been chosen to cover a range of possibilities for using networks in biological investigations. The first example uses a network to show that a phylogenetic pattern can actually be simpler than it appears from a tree-based analysis. The second example illustrates the analysis of complex patterns using networks, showing the limitations of the methodology. The third example is used to test an explicit hypothesis about evolutionary history. The fourth example uses a network as an exploratory tool, leading to explicit hypotheses about evolutionary events that can be subject to subsequent independent tests.

I have restricted the examples to the use of character-display methods, mainly because of the availability of suitable computer software. I have, thus, made no a priori assumptions about the possible biological causes of the apparent reticulations observed in the networks. Instead, I have merely approached each data set with the idea that studies of within-species variation should always take into account the possibility that reticulations could exist, and that uncovering them may provide important insights into the phylogenetic history of the organisms that will not be revealed by shoe-horning the data into a strictly tree-like model. In the process, I think that I may have found out more about the data than the original authors did, since a network analysis is likely to reveal more about the data than will a tree-based analysis. However, in most cases I have deliberately avoided detailed discussion of the biological conclusions arising from the analyses, and have instead focused on the analytical methods.

4.1. Sarcoptes mites

This first example represents a fairly simple application of networks. Skerrat et al. (2002) have reported partial sequence data for the mitochondrial small-subunit (SSU) ribosomal RNA gene for 23 specimens, representing nine haplotypes, of the parasitic mite *Sarcoptes scabiei* (Acari: Sarcoptidae) from wombat, human and dog hosts in Australia. They analysed these data phylogenetically to test scenarios concerning the origin and diversification of the scabies infections in wombats. The data were analysed using the maximum parsimony criterion in order to reconstruct the phylogeny of the haplotypes. They found multiple equally parsimonious trees, of which they presented one. Clearly there is either character conflict or ambiguity in the data, otherwise there would be only one most-parsimonious tree.

Furthermore, the parsimony analysis assumes that the underlying model for the evolutionary history of the organisms is a dichotomously branching tree. It is highly questionable whether this is an appropriate model for the relationships between individuals within a species. If nothing else, the possibility of reticulate evolution should be explicitly considered. So, a network analysis would be a better starting point. If there is no conflict among the characters then the network will form a dichotomous tree, but otherwise there will be reticulation among the branches, and the amount of reticulation will reflect the degree to which the data are or are not tree-like. It is also possible that the reticulate diagram might represent the true phylogenetic history of the mites.

So, I have added to the data set a sequence from a mite caught on a Swedish fox, and re-analysed the data. As an illustration of a phylogenetic tree, I performed a neighbour-joining analysis based on the Hasegawa–Kishino–Yano (HKY) substitution model, using the T-Rex v1.2a4 program (Makarenkov, 2001). The unrooted tree is shown in Fig. 1a. Note that the bootstrap values are very small (only two of them exceed 50%, for example), which is indicative of the degree of character conflict in the data, which also leads to the multiple parsimony trees. Indeed, the poor support for all of the branches suggests that the data might actually be of little value for reconstructing a phylogeny.

Nevertheless, we could proceed to interpret the tree as a phylogeny by providing a root. This particular tree could be rooted in either of two ways. First, the Fox sequence could be used as an outgroup, since it did not come from the same continent as the other sequences. Second, coalescent theory in population genetics suggests that the most likely root of any population-level phylogeny will be the most common (i.e. with the greatest number of representatives) or the most connected (i.e. with the most branches connected to it) of the haplotypes, which in this case is haplotype One (as it has 10 representatives and three branches). Note, also, that several of the haplotypes do not appear on pendant branches, but are instead ancestral haplotypes that have been sampled in the data set, which is quite a common occurrence for population-level studies.

An alternative analysis strategy, which I am emphasising in this paper, is to explore the data more thoroughly using a network analysis. In this case we need a character-display network, so that we can visualise the location and nature of the apparent conflict in the data. It turns out not to matter overly much which analysis is performed for this data set. I tried: (i) a median network, using the Spectronet v1.2 program (Huber et al., 2002); (ii) split decomposition with both the HKY substitution model and parsimony splits, using the SplitsTree v3.2 program (Huson, 1998); (iii) neighbour-net, using the SplitsTree v4b10 program (Huson and Bryant, 2005); and (iv) statistical parsimony, using the TCS v1.13 program (Clement et al., 2000). The unrooted network is shown in Fig. 1b. All of the analyses produced this network except for statistical parsimony, which produced a network that did not contain the node marked with an asterisk.

In this example, the unrooted network (Fig. 1b) is a complete visual representation of the data; that is, it shows all of the character-state patterns (which is not always true for a network, if the patterns are too complex to be shown in two dimensions). There are eleven sets of edges (branches) in the network, with six single edges (numbered 1–6) and five sets of multiple edges (numbered 7–11). These represent the eleven patterns of nucleotide variation observed among the haplotypes, with two sets of edges (6 and 10) having double length because those patterns occur twice in the data set. The distances along the edges between the haplotypes, thus, represent the nucleotide distances between them, each single-length branch requiring one nucleotide change. Each path through the network is supported by one set of compatible nucleotide characters, and if there is more than one path connecting two haplotypes then there are several (incompatible) sets of characters indicating alternative relationships among those haplotypes.

This diagram makes the incompatibility problem obvious. There is, in fact, very little conflict in the data at all, but one of the characters conflicts with three of the other characters. This character creates a split with haplotypes Eight + Two + Three in one group and the rest of the haplotypes in the other group, which is represented by the set of four vertical edges shown in Fig. 1b with label 9. That is, this character has the same character state in haplotypes Eight, Two and Three and a different character state in the other haplotypes. This pattern contradicts the patterns shown by the other characters (edges labelled 7, 10 and 11), and this creates three of the four boxes shown in the network. Without this character, there would only be one reticulation in the network (created by edges 7 and 8, which also represent conflicting character patterns), and so it would be almost tree-like. The tree shown in Fig. 1a is thus misleading, as it indicates poor support for all of the branches but does not show that this lack of support is created by only one conflicting character.

We could now proceed to try to interpret the network as a phylogeny by providing a root, as above. This is shown in Fig. 1c. Note that the two alternative root positions do not affect most of the branch directions, and thus the inferred phylogenetic history, but three of the branches (marked with a cross) would have their directions reversed. However, any such rooting procedure would probably be at least partly inappropriate in this case, because it is unlikely that all of the network nodes represent real ancestors, the way that they do in a tree. Indeed, it is quite likely that the reticulation node missing from the statistical parsimony network does not represent an ancestor but is merely there to help display the character conflict in the other networks. Each of the other reticulation nodes would also need to be evaluated individually, in a similar manner, in order to decide how best to interpret them. However, the data are not really sufficient on their own for reconstructing the history of these haplotypes, and thus any more detailed interpretation would probably be premature.

4.2. *Dictyocaulus lungworms*

I chose this example to represent the other extreme of complexity in network analyses. Höglund et al. (2004) have reported AFLP data for 72 specimens, representing nine isolates, of the lungworm *Dictyocaulus viviparus* (Nematoda: Dictyocaulidae) from cattle hosts in Sweden plus a laboratory strain. They analysed these data phylogenetically to examine the population genetic structure, using the distance-based neighbour-joining method.

The resulting tree diagram is shown in Fig. 2a. The character conflict is represented by the bootstrap percentages, six of which are large and three of which are small, and six of which are $<50\%$ and so not shown. This can be interpreted as indicating that the laboratory isolate and several of the farms (notably those with long branch lengths) form coherent genetic groups. There is a large amount of genetic variability within each farm, but the variability is structured so that the worms are still identifiable as coming from a particular farm. If the tree is rooted on the branch separating the laboratory strain from the field isolates then each farm forms a separate phylogenetic lineage.

However, the genetic pattern is not really that simple, because the tree-based analysis forces the data into a simple hierarchical structure by artificially ignoring the character conflict. This is what the small bootstrap values are indicating. A more realistic assessment of the data would involve a network analysis instead, which will try to display at least some of the conflicting patterns. In this case, it matters very much which network analysis is chosen, because the different methods respond differently to complex patterns such as exist in this data set.

I tried the same network analyses as used in the previous example, based on the distance matrix used to produce the neighbour-joining tree. The results of the neighbour-net analysis are shown in Fig. 2b. If you look carefully, you will note that the relationships among the nine isolates are the same as for the neighbour-joining tree but that each of the branches of that tree has been ‘thickened’ by the addition of reticulations. In other words, this looks more like a real living tree than the stick diagram in Fig. 2a. It is thereby a far more accurate representation of the character patterns in the data, because it emphasises the complexity in the pattern of shared character states among the samples. More to the point, it displays a more likely scenario of inter-relationships among populations due to inter-breeding, and creates a feeling of much less confidence in the hierarchical nature of the population genetic structure than does the tree diagram. In particular, there are clearly individuals from several of the farms that share character states in common with several of the individuals from other sources—for example, there

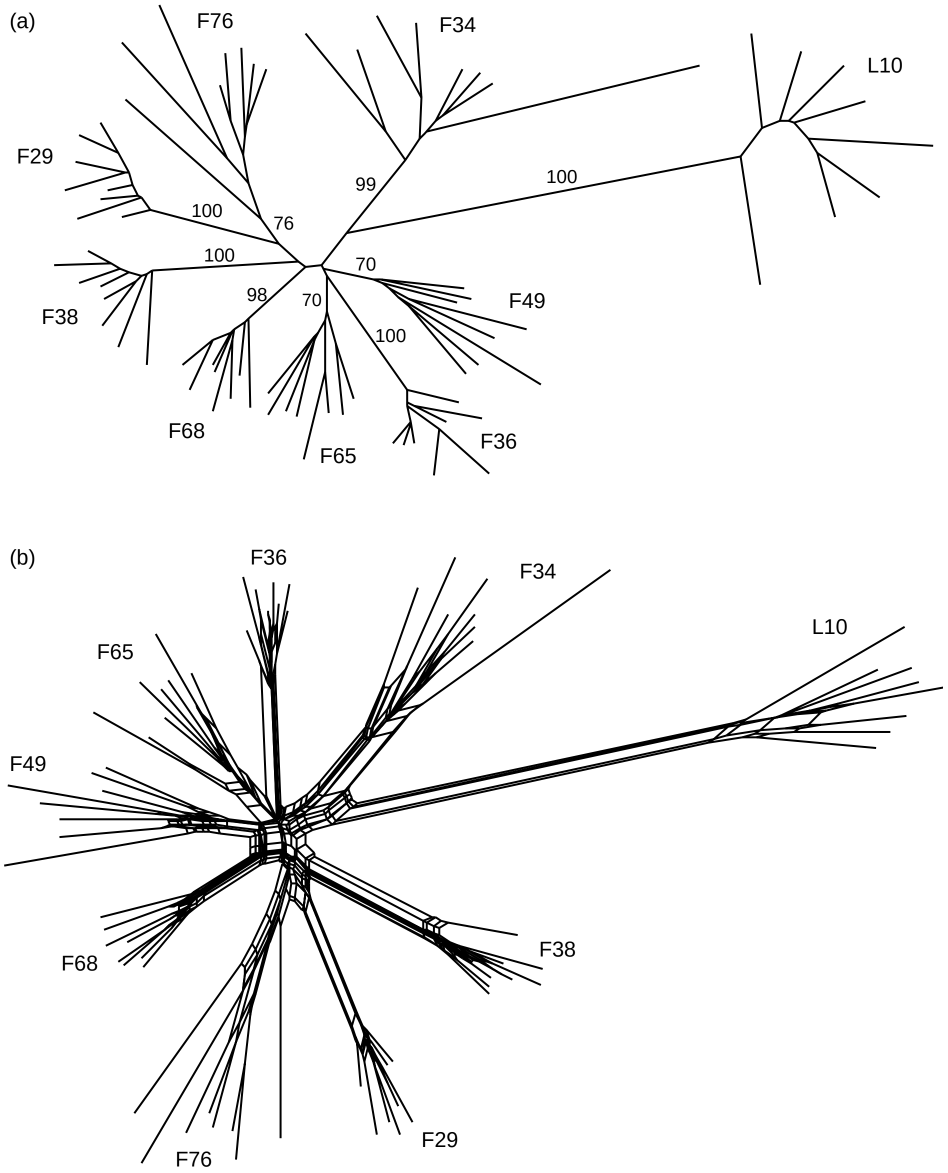

Fig. 2. Phylogenetic analyses of amplified fragment length polymorphisms for *Dictyocaulus viviparus* from cattle hosts on eight farms in Sweden (labelled F) plus a laboratory strain (labelled L). (a) Inferred unrooted tree from a neighbour-joining analysis of distance data. Each terminal branch represents one individual worm, and the numbers are bootstrap values as the percentage of 1000 replicates. (b) Inferred neighbour-net network from the same distance data matrix, displaying the conflicts among the characters as extra edges in the diagram.

are two worms from farm F65 that are linked to all of the worms on farms F49 and F68 while the other F65 worms are not, and there is one individual from farm F34 that is linked to the worms from the laboratory isolate. These patterns cannot be seen in the neighbour-joining tree.

The alternative network analyses have problems displaying the complex patterns shown in Fig. 2b. For example, the split decomposition analysis produces separate branches for each isolate but merely puts an unresolved bush at the end of each branch. The median network produces more than 100,000 nodes, which is undisplayable, so I tried producing a reduced median network instead. This revealed a series of blobs representing each of the isolates, but no details about relationships within or between the blobs could be resolved. The statistical parsimony analysis indicated that most of the differences between worms were beyond the parsimony connection limit, so it simply connected three of the worms from farm F36 and left every other worm in its own group. These diverse responses show that the various network methods respond differently to increasing data complexity.

It would also be possible to root the network in Fig. 2b using the laboratory isolate. However, this would clearly be unrealistic in this case, due to the complexity of the interrelationships. The diagram is meant to be a character display in this case, and it is best to leave it as such.

4.3. More *Sarcoptes* mites

This example is designed to illustrate the use of networks for testing explicit phylogenetic hypotheses. Walton et al. (2004) have reported partial sequence data for the mitochondrial SSU ribosomal RNA gene for 25 specimens, representing 14 haplotypes, of the parasitic mite *S. scabiei* (Acari: Sarcoptidae) from human hosts in Australia and Panama and from animal hosts of various locations. They analysed these data phylogenetically to test scenarios concerning the epidemiology of the scabies infections in humans. The data were analysed using neighbour-joining with the TN substitution model (and various other methods) in order to reconstruct the phylogeny of the haplotypes.

Walton et al. (2004) do not display the tree for the SSU data, but it is similar to the one they show for the cytochrome oxidase I gene sequences. More importantly, they do not specify a root for their trees. However, they do recognise three ‘groups’ in their data, and if these groups are interpreted in a phylogenetic sense (as they must be in order to reach several of the authors’ conclusions) then the inferred root must be somewhere on the three long branches in the middle of the diagram. The authors conclude that human–human transmission is the most important epidemiological pattern as far as control programs are concerned.

As always, when dealing with population-level data it is best to consider analyses that explicitly deal with reticulate phylogenetic history. Therefore, a network analysis might be more informative than a tree-based one for these data. I re-analysed the same data set using various distance-based network methods, based on the K3P substitution model. The split decomposition network is shown in Fig. 3. It is a simpler version of the network produced by the neighbour-net analysis, but shows the same essential features.

This network is basically the same as the tree diagram. That is, if the root of the diagram is in the middle, then the sequences from the Panama human mites form one group (on the left of the diagram), while the sequences from human mites in Australia form two groups, one of which is strongly associated with sequences from animal hosts. Epidemiologically, then, we can conclude that human–human transmission is important, because the mite sequences collected from humans form distinct genotype groups, and this may be taken as representing different groups of mites spreading from one person to another. Furthermore, we can also conclude that animal–human transmission is important, because several of the people have mites with genotypes that are common on all of the animals sampled, even animals that were sampled on different continents.

However, the network analysis goes beyond this simple interpretation, and provides extra information with regard to testing these epidemiological hypotheses. The series of boxes appearing along several of the branches indicate some character conflict, and in this case these probably represent evolutionary history. For example, the hominis8 + hominis9 sequences are shown as having a character pattern in common with the hominis200 + hominis205 sequences, in spite of the fact that these sequences were collected on different continents. This pattern cannot be seen in the tree diagram, and it is likely to be related to the source of the relatively recent introduction of scabies into Panama.

More importantly, however, there is a character pattern separating the dog-host and human-host sequences into two groups (indicated by the set of three edges marked with an asterisk in Fig. 3). This can be interpreted as representing two independent transmission routes between dogs and humans. More specifically, since the human-derived sequences are on pendant branches in both cases, the route of transmission is from the dogs to the humans rather than the other way around. Thus, we can conclude that animal to human transmission has occurred multiple times, and is thus an epidemiological phenomenon that cannot be ignored in a control program. This pattern also cannot be seen in the tree diagram.

These conclusions are clearly dependent on interpreting the network as a rooted representation of a phylogeny (i.e. a directed acyclic graph). This does not, however, mean that all of the nodes on the network necessarily represent unobserved ancestors. Moreover, it raises the point of where

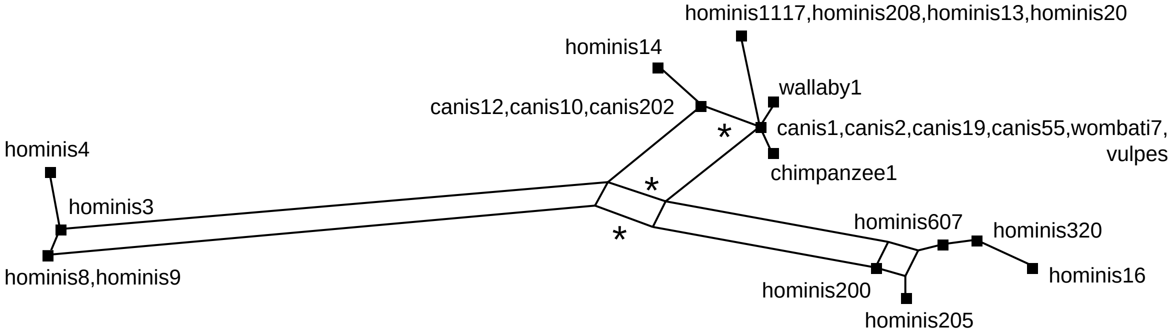

Fig. 3. Phylogenetic analyses of haplotype data for *Sarcoptes scabiei* from various hosts, based on mitochondrial SSU rRNA sequences. Each of the 14 haplotypes is marked as a box at the appropriate node, with the labels indicating the host from which the mite was collected (the numbers are merely codes for the individuals). The mites from dog hosts have been identified as var. *canis*, those from human hosts as var. *hominis* and those from wombat hosts as var. *wombati*. The asterisks indicate the set of three edges splitting the canis + hominis sequences into two groups.

the root should be placed. Coalescent theory in population genetics suggests that the root is not in the middle of the diagram at all, but should be placed at the node representing the haplotype for samples canis1 + canis2 + canis19 + canis55 + wombati7 + vulpes (as this is the both the most common and the most connected haplotype). This is consistent with the idea that the human scabies infections are ultimately derived from an animal source; and since the animals sampled are from several continents this is a worldwide set of genotypes. This change in root does not affect any of the conclusions listed above. However, it does mean that 'group C' of Walton et al. (2004) (i.e. the animal- + human-host group) is not actually an evolutionary group at all, but is the paraphyletic ancestral group from which the other two groups of derived human-host mites have descended.

The pattern shown by the cytochrome oxidase I gene sequences is similar to the one I have described here. However, the reticulations among the animal-associated genotypes are slightly more complex.

4.4. Toxoplasmatinae protozoans

This final example illustrates the use of networks in generating hypotheses that can be subsequently tested experimentally. The purpose here is to use the network analysis to provide a clear picture of the patterns in the data, so that explicit hypotheses can be generated that might explain those patterns, and then to briefly evaluate (i.e. informally test) those hypotheses. The example is based on analysis of the 19 available database sequences for the nuclear small-subunit (SSU) ribosomal RNA gene for *Hammondia hammondi*, *Neospora caninum* and *Toxoplasma gondii* (Apicomplexa: Sarcocystidae). The within-species variation of *T. gondii* among some of these sequences has previously been investigated (Luton et al., 1995), but with few notable conclusions. Note also that a study of within-species variation does not necessarily involve population-level processes, and thus this example extends the application of network analysis beyond the previous examples.

The database accession numbers for the sequences used are shown in Fig. 4, these being all of the available almost-complete sequences (the TGG sequence is taken from the release of the *T. gondii* genome-sequencing project current at the time of writing). I performed a neighbour-joining analysis based on the Kimura 2-parameter (K2P) substitution model, which produced the tree shown in Fig. 4a. This tree shows considerable variation among the sequences within *T. gondii*, but much less variation within *N. caninum*. Moreover, the topology seems to be uncontroversial. If the tree is rooted using the available sequences of *Besnoitia* (AF109678, AF291426, AY665399) then the ancestor is placed on the branch connecting the *Hammondia* and *Neospora* sequences, and we could interpret the tree as representing the phylogenetic history of the sequences.

However, what this tree does not reveal is that two of the branches represent exactly the same character-state changes within *Toxoplasma* and *Neospora*. These branches are marked with an asterisk in Fig. 4a. This fact cannot escape notice in a network analysis, but it is not immediately obvious either by looking at a tree or by scanning a sequence alignment. To demonstrate this, I performed a split decomposition analysis using the same data and model, as shown in Fig. 4b. The two boxes in the diagram are created by the fact that the character-state changes in question (also marked with asterisks in the network) form a contradictory pattern to the other characters, since this is a pattern shared by some of the sequences within each of the two species but not by the others.

Note that there would be little purpose in trying to turn this network into a directed phylogeny by adding a root, as was done for the tree. The network is displaying character conflict, and the nodes involved in the reticulations should not be interpreted as unobserved ancestors in this case. If these nodes are ignored, then the network is the same as the tree, and the tree is, thus, the best representation of the inferred phylogeny. Here, the network is merely being used as a valuable tool for investigating and displaying the character patterns.

There are two things worth noting about this intriguing character-state pattern. First, there are actually several characters involved in creating the pattern, at two different locations within the DNA sequence. The strongest support is at positions 1602–1603 (using the numbering system of Gagnon et al., 1996), where a 'GC' motif is inverted to 'CG'. All of the patterns are located within helices, which means that they can have an effect on the secondary structure of the rRNA molecule. Second, the pattern is created by multiple sequences from the same strains. That is, several strains have sequences on both sides of the split in the network for both *N. caninum* (NC-1) and *T. gondii* (RH, Me49). Thus, the pattern is not only a within-species pattern but is a replicated within-strain pattern.

This unexpected pattern is interesting, and requires some explanation. That is, it generates a set of possible explanatory hypotheses that can be further investigated. As described in Section 2.2, there are three possible explanations for a reticulation in a character-display network, and therefore three possible types of hypotheses that could explain the current pattern. First, it could be uncertainty in the data, such as sampling error. For example, given the relatively small number of DNA differences observed here, we should not exclude the possibility of sequencing errors. However, the probability of coincident errors in multiple strains across two species is quite small, unless there is some particular feature of the SSU sequence at these locations that make them particularly prone to sequencing error. Second, it could be homoplasy (analogy). In this case, this would involve coincident mutations in the ancestral sequences of *T. gondii* and *N. caninum*. Once again, this is possible but perhaps unlikely, unless no other

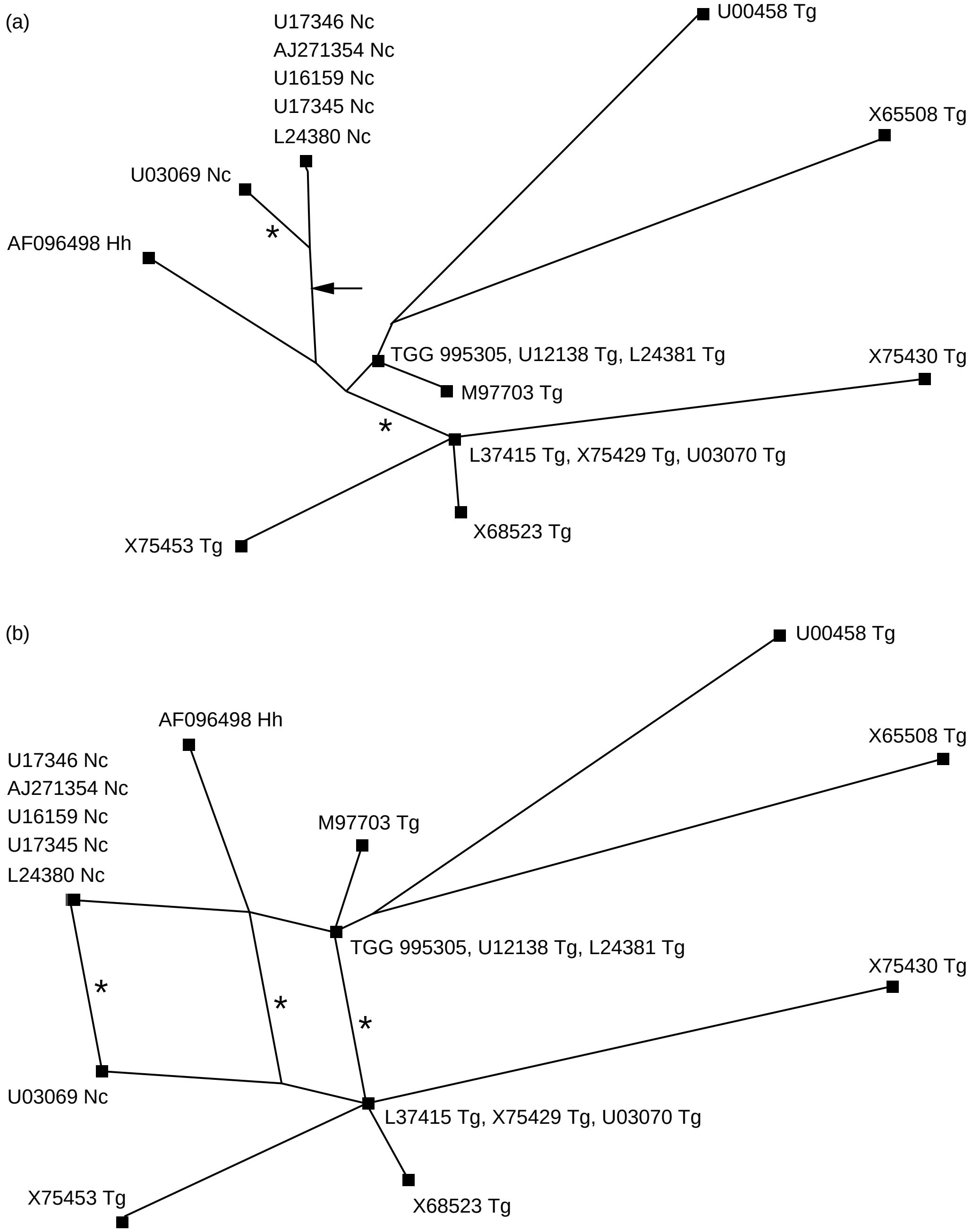

Fig. 4. Phylogenetic analyses of sequence data for *Hammondia hammondi* (Hh), *Neospora caninum* (Nc) and *Toxoplasma gondii* (Tg), based on nuclear SSU rRNA sequences. Each of the 19 sequences is labelled with its database accession number, and its position is marked with a box at the appropriate node. (a) Inferred unrooted tree from a neighbour-joining analysis using the K2P substitution model. Using the *Besnoitia* sequences as an outgroup roots the tree on the branch indicated by the arrow. The asterisks mark the two branches with the same set of character-state changes. (b) Inferred split decomposition network based on the same distance data. The asterisks mark the set of three parallel edges corresponding to the two asterisked branches in (a).

explanation presents itself. Third, the pattern could represent a single evolutionary event (i.e. homology). (It could also represent multiple homologous events in the same lineage, but that would be a less parsimonious explanation).

There are at least two possible such ancestral events that could lead to the current pattern. One is the idea of ancestral recombination. That is, some recombination event pre-dating the split of *Neospora* and *Toxoplasma* has left a partial signal in the SSU sequences, but this has been overlain by subsequent evolutionary events in other gene sequences. Formally testing for recombination in sequence data can be a tricky business, as there are many genetic signatures that could be left by recombination events, and each proposed method only tests for a small subset of these. So, I tested for recombination using the RDP v2beta08 program (Martin et al., 2005) and the RecombiTEST suite (Piganeau and Eyre-Walker, 2004), both of which implement a range of tests for recombination. Recombination was detected by four of the seven tests. So, this possibility cannot be excluded, and would perhaps be worth pursuing in more detail.

The second ancestral event that could have created the observed pattern is divergent paralogy pre-dating the split of *Neospora* and *Toxoplasma*. In plants and animals rDNA is usually arranged as thousands of copies (paralogues) of sequential units located in several arrays, and the copies are relatively homogeneous as a result of concerted evolution. However, in several members of the Apicomplexa it is now recognised that there are relatively few rDNA units and that these are divergent paralogues (i.e. at least some of the paralogues have different DNA sequences). These include members of *Cryptosporidium* (Le Blancq et al., 1997), *Plasmodium* (Rooney, 2004); *Babesia* (Dalrymple, 1990; Reddy et al., 1991), *Theileria* (Kibe et al., 1994), and most importantly *Neospora* (Gondim et al., 2004). In particular, for *N. caninum* the ITS1 sequence has been shown in several strains to contain two different DNA sequences, although this could not be shown for the examined strains of *Toxoplasma* (Gondim et al., 2004). The implication here is that the paralogy extends to the whole rDNA unit, and that it exists in both species, an hypothesis that has apparently not been proposed before. If this is so, then it has consequences for the use of any rDNA sequence in phylogenetic analyses of the Coccidia, as it is likely that different paralogues are amplified with different degrees of success and thus are more or less likely to be included in an analysis. The only exception to this reported pattern for the Apicomplexa is the suggestion by Guay et al. (1992) that *T. gondii* has ∼110 copies of the 5S gene, which is located within the rDNA unit. The forthcoming release of the final results from the *T. gondii* genome project will presumably provide a definitive test of this hypothesis.

As can be seen, the pattern observed in the network representation generates a set of hypotheses that are amenable to experimental test. The pattern would not have been observed if only a tree-based phylogenetic analysis was performed, and thus none of these hypotheses would be proposed or tested.

5. The future

Currently, there is no good general method for deriving networks, in the sense of an all-purpose tool for generating diagrams with reticulations. This is partly because there is a lack of reasonable evolutionary models for generating reticulations, comparable to the models that have been developed for dichotomous branching processes. So, perhaps a combination of methods will always be needed, designed and chosen for specific purposes.

For the ‘true’ phylogenetic networks, the key question is: can they be generalised to simultaneously include all possible causes of reticulation? This will require a decision about the ‘signatures’ in the data. If each cause of reticulation leaves a different data signature, then it is theoretically possible to devise models to detect those signatures and to present an optimal representation of how the signature might have arisen. Some suggestions are made, for example, by Linder and Rieseberg (2004) regarding recombination versus hybridisation. However, this could create a very complicated (and possibly intractable) situation mathematically.

For the display-type networks, the key question is: can they be rooted/directed consistently (i.e. turned into rooted directed acyclic graphs) so that they can be interpreted phylogenetically? What we would need is a method to remove branches/nodes that do not represent truly homologous phylogenetic events. This may be possible, but it goes well beyond what the methods were originally intended for.

There is also currently a lack of generally accepted criteria for assessing the quality of a network versus a tree. It is not straightforward to compare a reticulation model to a dichotomous model, in terms of goodness-of-fit to the data. While some progress has been made (Moret et al., 2004), the fact that trees are nested within networks makes the comparison problematical. Ultimately, this may be the biggest stumbling block, as biologists might be reluctant to go beyond simple tree models unless there is a clear optimality criterion for deciding when a more complicated model is necessary (Legendre, 2000a).

There will also need to be detailed comparative studies of the various network methods, to compare and contrast their success rates in terms of false negatives and false positives. These studies could involve simulated data (e.g. Nakhleh et al., 2003), real data resulting from known reticulation events in a phylogenetic history (e.g. McDade, 1997), or some combination of the two (e.g. real data re-arranged to create artificial recombination/hybridisation events).

References

Addario-Berry, L., Hallett, M., Lagergren, J., 2003. Towards identifying lateral gene transfer events. Pac. Symp. Biocomput. 8, 279–290.

Aude, J.-C., Diaz-Lazcoz, Y., Codani, J.-J., Risler, J.-L., 1999. Applications of the pyramidal clustering method to biological objects. Comput. Chem. 23, 303–315.

Bandelt, H.-J., 1994. Phylogenetic networks. Verhandl. Naturwiss. Vereins Hamburg 34, 51–71.

Bandelt, H.-J., Dress, A.W., 1989. Weak hierarchies associated with similarity measures: an additive clustering technique. Bull. Math. Biol. 51, 133–166.

Bandelt, H.-J., Dress, A.W.M., 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1, 242–252.

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

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

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

Bandelt, H.-J., Macauley, V., Richards, M., 2000. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Mol. Phylogenet. Evol. 16, 8–28.

Bryant, D., 2003. A classification of consensus methods for phylogenetics. In: Janowitz, M.F., Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (Eds.), Bioconsensus. American Mathematical Society Publications, Piscataway, NJ, pp. 163–183.

Bryant, D., Moulton, V., 2002. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. Lect. Notes Comput. Sci. 2452, 375–391.

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

Castelloe, J., Templeton, A.R., 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogenet. Evol. 3, 102–113.

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

Dalrymple, B.P., 1990. Cloning and characterization of the rRNA genes and flanking regions from *Babesia bovis*: use of genes as strain discriminating probes. Mol. Biochem. Parasitol. 43, 117–124.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.

Doolittle, W.F., 1999. Phylogenetic classification and the universal tree. Science 284, 2124–2128.

Eigen, M., Winkler-Oswatitsch, R., Dress, A., 1988. Statistical geometry in sequence space. Proc. Natl Acad. Sci. USA 85, 5913–5917.

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

Fitch, W.M., 1997. Networks and viral evolution. J. Mol. Evol. 44 (Suppl.), S65–S75.

Fluxus Technology 2000. Network: A Program Package for Calculating Phylogenetic Networks, program and documentation, Fluxus Engineering, Berlin.

Gagnon, S., Bourbeau, D., Levesque, R.C., 1996. Secondary structures and features of the 18S, 5.8S and 26S ribosomal RNAs from the Apicomplexan parasite *Toxoplasma gondii*. Gene 173, 129–135.

Gondim, L.F.P., Laski, P., Gao, L., McAllister, M.M., 2004. Variation of the internal transcribed spacer 1 sequence within individual strains and among different strains of *Neospora caninum*. J. Parasitol. 90, 119–122.

Guay, J.-M., Huot, A., Gagnon, S., Tremblay, A., Levesque, R.C., 1992. Physical and genetic mapping of cloned ribosomal DNA from *Toxoplasma gondii*: primary and secondary structure of the 5S gene. Gene 114, 165–171.

Gusfield, D., Eddhu, S., Langley, C., 2004. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. J. Bioinform. Comput. Biol. 2, 173–214.

Hallett, M., Lagergren, J., 2001. Efficient algorithms for lateral gene transfer problems. In: Lengauer, T. (Ed.), Proceedings of the Fifth Annual International Conference on Research in Computational Molecular Biology. ACM Press, New York, pp. 141–148.

Hallett, M., Lagergren, J., Tofigh, A., 2004. Simultaneous identification of duplications and lateral transfers. In: Bourne, P.E., Gusfield, D. (Eds.), Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology. ACM Press, New York, pp. 347–356.

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

Hein, J., 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. 36, 396–405.

Höglund, J., Engström, A., Morrison, D.A., Mattsson, J.G., 2004. Genetic diversity assessed by AFLP analysis of the parasitic nematode *Dictyocaulus viviparus*, the lungworm of cattle. Int. J. Parasitol. 34, 475–484.

Holland, B., Moulton, V., 2003. Consensus networks: a method for visualizing incompatibilities in collections of trees. Lect. Notes Comput. Sci. 2812, 165–176.

Huber, K.T., Moulton, V., Lockhart, P., Dress, A., 2001a. Pruned median networks: a technique for reducing the complexity of median networks. Mol. Phylogenet. Evol. 19, 302–310.

Huber, K.T., Watson, E.E., Hendy, M.D., 2001b. An algorithm for constructing local regions in a phylogenetic network. Mol. Phylogenet. Evol. 19, 1–8.

Huber, K.T., Langton, M., Penny, D., Moulton, V., Hendy, M., 2002. Spectronet: a package for computing spectra and median networks. Appl. Bioinform. 1, 159–161.

Huson, D.H., 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68–73.

Huson, D.H., Bryant, D., 2005. Estimating phylogenetic trees and networks using SplitsTree4, program and documentation. Algorithms in Bioinformatics, University of Tübingen, Tübingen.

Kibe, M.K., ole-MoiYoi, O.K., Nene, V., Khan, B., Allsopp, B.A., Collins, N.E., Morzaria, S.P., Gobright, E.I., Bishop, R.P., 1994. Evidence for two single copy units in *Theileria parva* ribosomal RNA genes. Mol. Biochem. Parasitol. 66, 249–259.

Lapointe, F.-J., 2000. How to account for reticulation events in phylogenetic analysis: a comparison of distance-based methods. J. Classific. 17, 175–184.

Le Blancq, S.M., Khramtsov, N.V., Zamani, F., Upton, S.J., Wu, T.W., 1997. Ribosomal RNA gene organization in *Cryptosporidium parvum*. Mol. Biochem. Parasitol. 90, 463–478.

Legendre, P., 2000a. Reticulate evolution: from bacteria to philosopher. J. Classific. 17, 153–157.

Legendre, P., 2000b. Biological applications of reticulation analysis. J. Classific. 17, 191–195.

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

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

Luton, K., Gleeson, M., Johnson, A.M., 1995. rRNA gene sequence heterogeneity among *Toxoplasma gondii* strains. Parasitol. Res. 81, 310–315.

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

Makarenkov, V., Legendre, P., 2004. From a phylogenetic tree to a reticulated network. J. Comput. Biol. 11, 195–212.

Martin, D., Williamson, C., Posada, D., 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21, 260–262.

McDade, L.A., 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. Evolution 46, 1329–1346.

McDade, L.A., 1997. Hybrids and phylogenetic systematics III. Comparison with distance methods. Syst. Bot. 22, 669–683.

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

Morrison, D.A., Ljunggren, E.L., Mattsson, J.G., 2003. The origin of *Sarcoptes scabiei* in wombats. Parasitol. Res. 91, 497–499.

Nakhleh, L., Sun, J., Warnow, T., Linder, C.R., Moret, B.M.E., Tholse, A., 2003. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. Pac. Symp. Biocomput. 8, 315–326.

Nakhleh, L., Warnow, T., Linder, C.R., 2004. Reconstructing reticulate evolution in species—theory and practice. In: Bourne, P.E., Gusfield, D.

(Eds.), Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology. ACM Press, New York, pp. 337–346.

Piganeau, G., Eyre-Walker, A., 2004. A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. Heredity 92, 282–288.

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

Reddy, G.R., Chakrabarti, D., Yowell, C.A., Dame, J.B., 1991. Sequence microheterogeneity of the three small subunit ribosomal RNA genes of *Babesia bigemina*: expression in erythrocyte culture. Nucleic Acids Res. 19, 3641–3645.

Rooney, A.P., 2004. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in Apicomplexans. Mol. Biol. Evol. 21, 1704–1711.

Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analyses. Genetics 156, 879–891.

Skerrat, L.F., Campbell, N.J.H., Murrell, A., Walton, S., Kemp, D., Barker, S.C., 2002. The mitochondrial 12S gene is a suitable marker of populations of *Sarcoptes scabiei* from wombats, dogs and humans in Australia. Parasitol. Res. 88, 376–379.

Smouse, P.E., 2000. Reticulation inside the species boundary. J. Classific. 17, 165–173.

Sneath, P.H.A., 1975. Cladistic representation of reticulate evolution. Syst. Zool. 24, 360–368.

Sneath, P.H.A., 2000. Reticulate evolution in bacteria and other organisms: how can we study it? J. Classific. 17, 159–163.

Stevens, P.F., 1994. The Development of Biological Systematics: Antoine-Laurent de Jussieu, Nature, and the Natural System. Columbia University Press, New York.

Strimmer, K., Moulton, V., 2000. Likelihood analysis of phylogenetic networks using directed graphical methods. Mol. Biol. Evol. 17, 875–881.

Strimmer, K., Wiuf, C., Moulton, V., 2001. Recombination analysis using directed graphical models. Mol. Biol. Evol. 18, 97–99.

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

von Haeseler, A., Churchill, G.A., 1993. Network models for sequence evolution. J. Mol. Evol. 37, 77–85.

Walton, S.F., Dougall, A., Pizzutto, S., Holt, D., Taplin, D., Arlian, L.G., Morgan, M., Currie, B.J., Kemp, D.J., 2004. Genetic epidemiology of *Sarcoptes scabiei* (Acari: Sarcoptidae) in northern Australia. Int. J. Parasitol. 34, 839–849.

Wang, L., Zhang, K., Zhang, L., 2001. Perfect phylogenetic networks with recombination. J. Comput. Biol. 8, 69–78.

Winkworth, R.C., Bryant, D., Lockhart, P.J., Havell, D., Moulton, V. Biogeographic interpretation of split graphs: least squares optimization of branch lengths. Syst. Biol. (in press).

Xu, S., 2000. Phylogenetic analysis under reticulate evolution. Mol. Biol. Evol. 17, 897–907.