MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 3385–3390

NEWS AND VIEWS

OPINION

Trees and/or networks to display intraspecific DNA sequence variation?

PATRICK MARDULYN

Evolutionary Biology and Ecology, Université Libre de Bruxelles, CP 160/12, av FD Roosevelt 50, 1050 Brussels, Belgium; Fonds de la Recherche Scientifique – FNRS, Brussels, Belgium

Abstract

Phylogenetic trees and networks are both used in the scientific literature to display DNA sequence variation at the intraspecific level. Should we rather use trees or networks? I argue that the process of inferring the most parsimonious genealogical relationships among a set of DNA sequences should be dissociated from the problem of displaying this information in a graph. A network graph is probably more appropriate than a strict consensus tree if many alternative, equally most parsimonious, genealogies are to be included. Within the maximum parsimony framework, current phylogenetic inference and network-building algorithms are both unable to guarantee the finding of all most parsimonious (MP) connections. In fact, each approach can find MP connections that the other does not. Although it should be possible to improve at least the maximum parsimony approach, current implementations of these algorithms are such that it is advisable to use both approaches to increase the probability of finding all possible MP connections among a set of DNA sequences.

Keywords: haplotype networks, intraspecific DNA sequences, parsimony, phylogenetic trees, phylogeography

Received 23 December 2011; revision received 12 March 2012; accepted 11 April 2012

Studies exploring DNA sequence variation at the intraspecific level commonly use two types of graph to summarize the genetic data: phylogenetic trees and/or haplotype networks. For example, a quick survey of phylogeographic studies published in the scientific journal *Molecular Ecology* in 2011 reveals that among 45 studies, 38% displayed DNA sequence variation using phylogenetic trees, 22% using haplotype networks and 40% using both trees and networks. Phylogenetic trees are connected graphs with no cycle (Huson *et al.* 2011) and are traditionally used to present estimates of phylogenetic relationships among species.

Correspondence: Patrick Mardulyn, Fax: 32 2 6502445; E-mail: pmarduly@ulb.ac.be

Phylogenetic networks are connected graphs with cycles. Although different types of network exist, I restrict the discussion here to what is commonly called haplotype or allele networks, generated by, for example, a median-joining (Bandelt *et al.* 1999) or statistical parsimony (Clement *et al.* 2000) analysis, in which nodes represent different allelic sequences, joined by edges (branches) whose length is defined and shows the number of nucleotides that differ between them (e.g. Huson *et al.* 2011). Split networks (Huson & Bryant 2006), built by combining each split (i.e. a partition of the sequences in two subsets) identified by the data, from either distances (Bandelt & Dress 1992; Bryant & Moulton 2004) or trees (Holland & Moulton 2003; Holland *et al.* 2004), also offer an interesting tool to explore sequence data and are increasingly used to display intraspecific sequence variation (e.g. Cassens *et al.* 2003; Marshall *et al.* 2009; Barrett & Freudenstein 2011). Their main purpose is to visualize the ambiguous phylogenetic signal present in a data set. They will, however, not be discussed here further because, as explained below, and unlike haplotype networks, they are not appropriate to summarize the information contained in several most parsimonious phylograms, which is the focus of this article.

Trees or networks?

Should we favour the use of trees and/or the use of networks? Inferring a genealogy from a set of intraspecific DNA sequences can be done using (i) phylogenetic tree methods, that is, classic methods of phylogeny inference that use an optimality criterion to compare trees (e.g. maximum parsimony and maximum likelihood) or (ii) network methods whose algorithms directly generate a network graph from the sequences (e.g. median-joining or statistical parsimony). While a network graph is typically tied to, and sometimes defined by, the algorithm that generates it, I suggest that to compare the pros and cons of trees and networks, it is useful to differentiate the algorithm used to infer the genealogical relationships among a sample of DNA sequences (i.e. the algorithm that connects the sequences together), from the graph used to display this information. The processes of (i) identifying the evolutionary (e.g. most parsimonious) connections among the sequences and (ii) drawing the graph displaying this information can be treated in principle independently from each other. In fact, under this rationale, Cassens *et al.* (2005) have proposed a method to create a network graph from the set of all most parsimonious trees, thereby combining a classic phylogeny inference method with an algorithm that constructs a network graph. Also, other algorithms have been designed previously for the purpose of inferring network graphs that include all most parsimonious trees,

directly from the sequence data: the median network algorithm of Bandelt *et al*. (1995), restricted to the analysis of strictly binary data (two character states), and an algorithm developed by Fitch (1997), which in practice is restricted to a limited number of sequences. If we do adopt a parsimony framework, which is reasonable when dealing with closely related sequences (e.g. Holder & Lewis 2003; Felsenstein 2004), as intraspecific sequences often are, we can rephrase the problem as follows: Which algorithm is best to infer all MP connections among a group of sequences, and which kind of graph is best to display them?

Tree graph or network graph?

While there are conceptual differences between a graph displaying genetic variation at the intraspecific level and a tree delivering an estimate of a species phylogeny (Posada & Crandall 2001), a tree graph can technically be used to display intraspecific DNA sequence variation. In fact, in a parsimony framework, a clear correspondence can be established between a strictly bifurcating phylogram (i.e. a tree whose branches lengths are proportional to the number of mutations that have occurred along them, as opposed to a cladogram that does not include branch length information) and a haplotype network, as already discussed in Cassens *et al*. (2005) and shown in Fig. 1. Indeed, both types of graph display the same information: the genetic distances separating alleles, the frequency of each allele (as displayed by the size of each circle in Fig. 1b and as the number of sequences in Fig. 1c) and even some information over their geographic distribution (using pie charts in networks or identifying the geographic location of each sequence in phylograms). Within a parsimony framework, a haplotype network without cycle is thus equivalent to a phylogram. Note that in this context, the length of a branch separating two alleles in both types of graph simply corresponds to the number of sites that are different between the two sequences, which is probably a good estimation of the number of mutations separating both sequences from their common ancestor, for closely related sequences.

Ambiguous connections appear under the parsimony criterion when two or more alternative connections share the same length, that is, in the presence of homoplasy in the sequence data. One can deal with ambiguity, as is usually done in phylogenetic studies, by considering all equally parsimonious trees separately, which rapidly becomes impractical with increasing numbers of trees, or by building a strict consensus cladogram compatible with all most parsimonious (MP) trees, that is collapsing clades, thereby creating a tree that includes multifurcations. But this has the undesirable effect of decreasing the amount of information contained in the graph (Fig. 2d), because (i) the resulting strict consensus tree can become compatible with many more trees than the initial set of MP trees that was used to build it and (ii) ambiguous information over branch length cannot be displayed. Another way of dealing with ambiguities is to draw a haplotype network that includes cycles

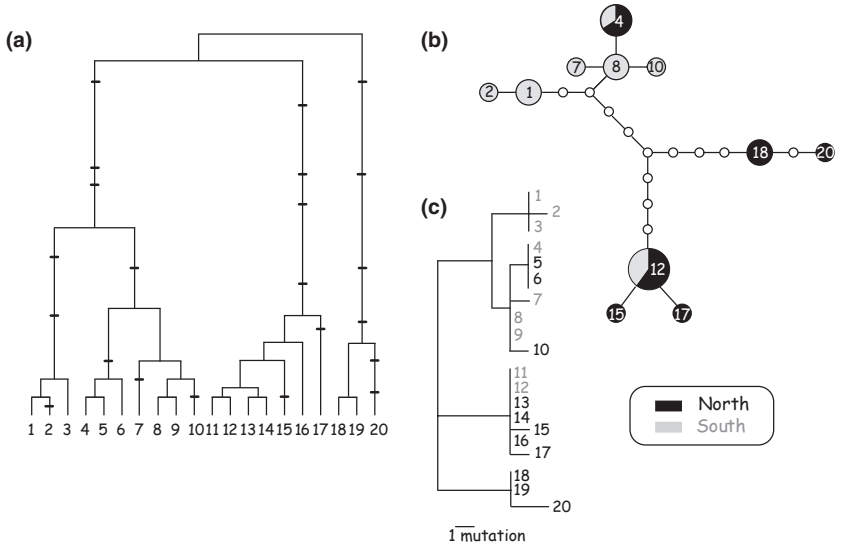

Fig. 1 Relationship between a gene genealogy, an estimated network graph and an estimated phylogram. (a) Gene genealogy displaying genealogical relationships among 20 sampled copies of a gene within a species. Branch length is proportional to time. Horizontal bars identify mutations along the branches. (b), Allele network showing the history of mutations that have occurred among the different alleles of the sampled sequences. Each allele is represented by a circle: its size is proportional to its frequency and its colour provides information about its geographic distribution. Each edge corresponds to a single mutation, and small white circles correspond to inferred alleles, absent from the data set. (c) phylogram showing the history of mutations that have occurred among the different alleles of the sampled sequences. (b) and (c) are equivalent in terms of information content and can easily be drawn from (a). However, (a) is usually unknown, and (b) and (c) are estimated from a set of sampled sequences.

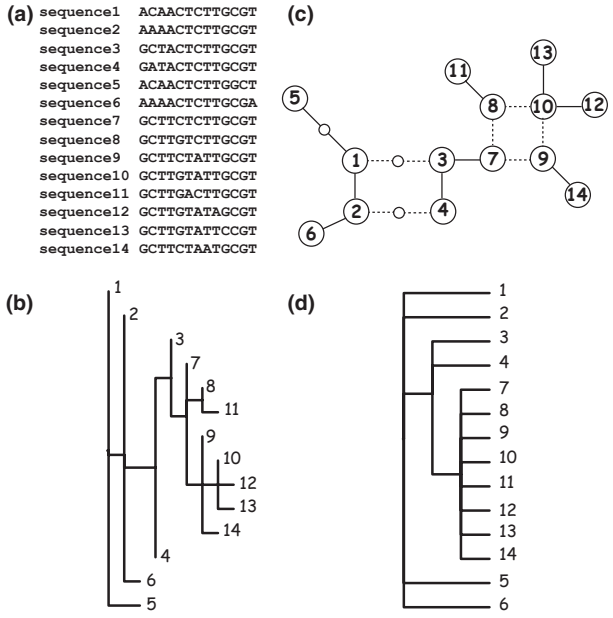

Fig. 2 Comparison of the informativeness of networks and strict consensus trees. (a) Polymorphic sites for a set of 14 sequences. (b) One of several most parsimonious phylograms associated with this set of sequences. (c) Network graph including all most parsimonious trees associated with this set of sequences. Each sequence in the data set is represented by a labelled circle; each branch represents one mutation; unlabelled circles are sequences not present in the data set. Dashed branches are those that may be deleted to recreate one of the MP phylogram used to create the graph. (d) Strict consensus cladogram of all most parsimonious trees associated with this set of sequences, as generated by PAUP* (Branch and Bound search).

(also called loops; Fig. 2c). The main feature distinguishing networks from trees is indeed the possibility of including cycles to represent (i) reticulate evolution (e.g. recombinations, hybridisations, lateral gene transfers) or (ii) conflicting signal in the data. Morrison (2005) makes a clear distinction between a 'true phylogenetic network' that aims to display true reticulate evolutionary events, and a 'character-display network', a network graph that merely displays all connections among the sequences that are equally well supported, revealing conflicting genealogical signal. While the ability to represent reticulate evolution is an essential feature of network graphs, the discussion here will be limited to the second type of network that summarizes all ambiguous connections in a single figure. A haplotype network with cycles can also be compatible with more trees than the initial set of MP trees, although the number of compatible trees is usually smaller than with a strict consensus tree. Moreover, this number can be further decreased by identifying branches (edges) that may be deleted in each cycle, and others that cannot, to recreate all the individual trees compatible with the network (Fig. 2c). In a parsimony framework, branches that can be discarded are those whose deletion results in a minimum length

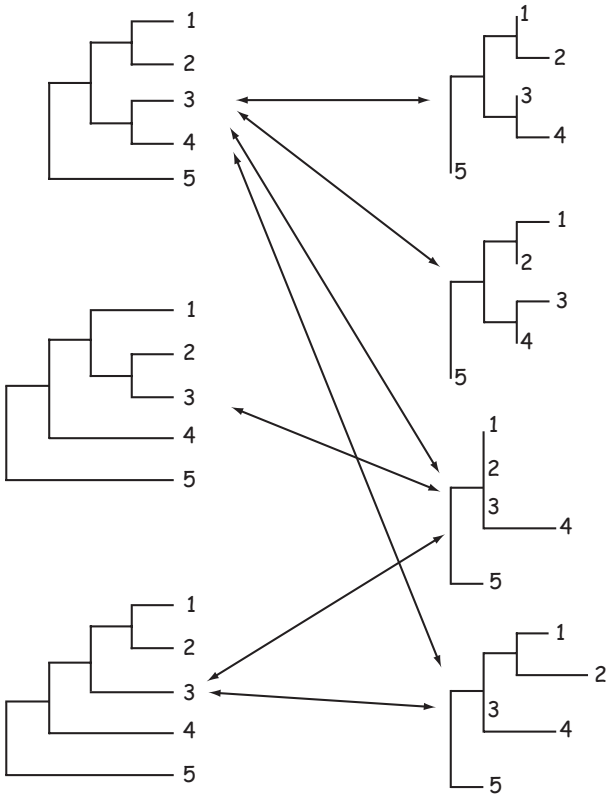

Fig. 3 Arrows indicate compatibility between three example MP cladograms and four example MP phylograms. A cladogram can be translated into a phylogram by assigning one possible MP ancestral sequence to each interior node of the tree. For one cladogram, alternative MP ancestral sequences may generate different phylograms. Conversely, some phylograms are compatible with more than one cladogram. A cladogram is considered compatible with a phylogram if it can be transformed into this phylogram only by modifying its branch lengths, including assigning a length of zero to some branches.

graph. Also, a network graph with cycles continues to display branch length information.

It should be noted that a split network cannot be used for the same purpose of displaying several MP phylograms into a single figure. A split network inferred from a set of MP trees (i.e. a consensus network; Holland & Moulton 2003) is built by combining all the splits defined by this set of trees and loses the information of branch lengths contained in the initial phylograms.

In conclusion, when alternative connections are equally parsimonious, a haplotype network graph appears more appropriate to summarize intraspecific DNA sequence variation, because it conveys more information than a strict consensus cladogram.

Phylogenetic estimation algorithm or network-building algorithm?

This does not mean, however, that network-building methods are automatically better. As argued by Cassens *et al.*

(2005), several methods of phylogenetic inference are based on the use of an optimality criterion to explore the space of all possible trees, which confers an advantage to them over network construction methods that are entirely defined by the algorithm used to construct the network step by step, and do not consider an a posteriori comparison of the resulting network to alternative networks based on some kind of criterion. Moreover, recent comparisons of the performances of traditional phylogenetic algorithms with those from network construction methods appear to be in favour of the former. Woolley *et al.* (2008) have tested the performances of these methods by analysing DNA sequence data simulated under various parameter values of a classic coalescent model depicting the evolution of a panmictic population. They concluded that although most methods perform equally well when the DNA sequence substitution rate is low, maximum parsimony is more accurate than network construction algorithms at a higher rate of substitution. Salzburger *et al.* (2011) have tested these performances under a wider range of population models, featuring coalescent simulations under both a single panmictic population and a structured population with symmetric or asymmetric migration. They concluded that (i) all traditional phylogenetic methods outperform the statistical parsimony network construction algorithm implemented in the software TCS (Clement *et al.* 2000), and (ii) in general, maximum parsimony analyses performed slightly better than maximum likelihood.

In the light of all of the above, it would seem logical to advise the use of the traditional maximum parsimony method to analyse intraspecific DNA sequences, and to combine all resulting MP trees in a network graph, using, for example, the algorithm UMP (Union of Maximum Parsimonious trees, Cassens *et al.* 2005). However, it is relatively easy to identify a set of sequences for which a median-joining analysis will perform better than maximum parsimony. For example, Mardulyn *et al.* (2009) have described a simple example data set (only four sequences of three nucleotides long: AAA, ACA, GAT, GCT) for which the median-joining or TCS algorithms performed better than maximum parsimony: a maximum parsimony PAUP* (Swofford 2003) search resulted in a single MP

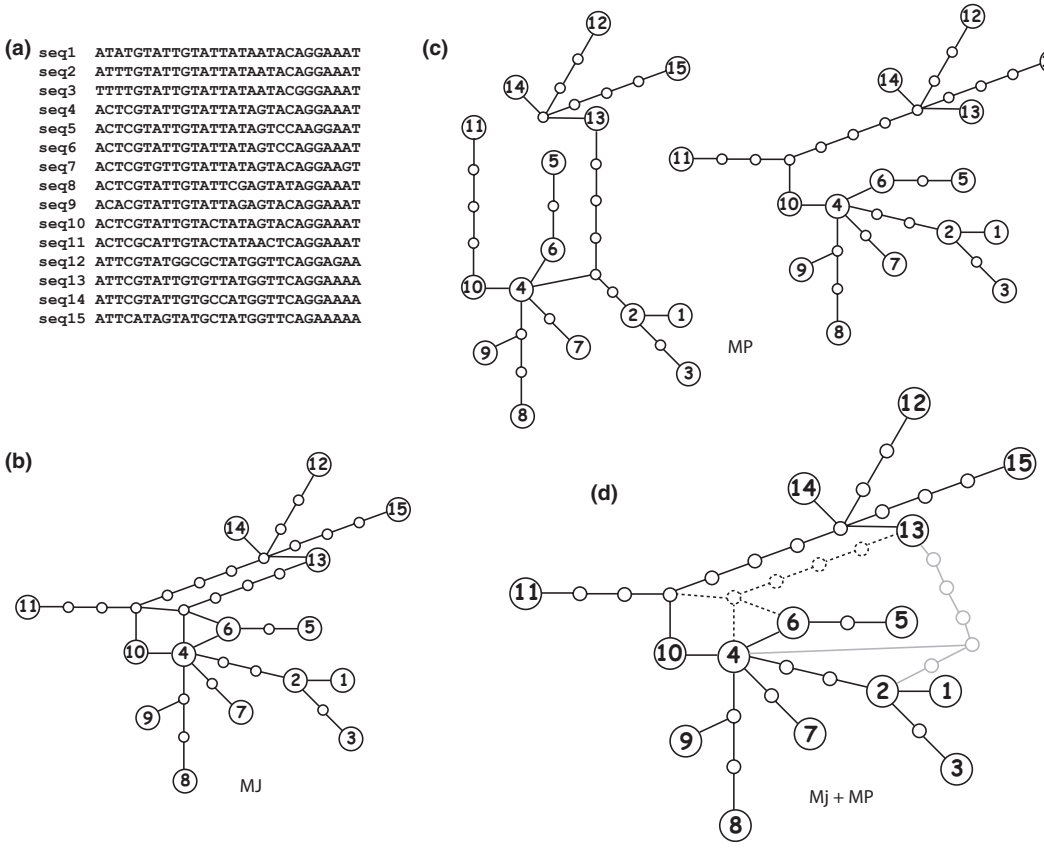

Fig. 4 Maximum parsimony and median-joining algorithms can both lead to different and unique most parsimonious solutions. (a) Polymorphic sites for a set of 15 DNA sequences. (b) A median-joining network inferred from the set of sequences in A using the program network (Bandelt *et al.* 1999; available at http://www.fluxus-engineering.com/sharenet.htm). (c) The two most parsimonious trees inferred from a Branch and Bound maximum parsimony search performed on the set of sequences in (a) using PAUP*. (d) Graph including all most parsimonious trees found in (b) and (c). Branches found in both (b) and (c) are in black. Branches unique to (b) are in grey. Dashed-lines branches are unique to (c).

phylogram, while the median-joining or TCS graph contained two MP phylograms, including the one found by the PAUP* analysis. The reason for the better performance of the network-building algorithms in this example lies with the way PAUP* deals with most parsimonious ancestral states. Like most phylogeny estimation programs, when running a maximum parsimony search, it compares cladograms (not phylograms) using the parsimony criterion. This is fine if one is interested in estimating evolutionary relationships among well-differentiated species, but a phylogram is more informative when considering intraspecific DNA sequence variation (Fig. 3 shows the compatibility relationships existing between some phylograms and cladograms). As already noted by Salzburger *et al.* (2011), more than one phylogram (called a Fitch tree in their article) corresponds to a single MP cladogram, because several equally parsimonious reconstructions of ancestral sequences are usually possible. If the inference program is subsequently asked to estimate branch lengths for the inferred set of MP cladograms, to generate a set of MP phylograms, it will assign only one of several possible most parsimonious ancestral sequences for each interior node of a tree. Because all alternative reconstructions of ancestral states are not taken into account, some MP phylograms may not be considered. The generated set of MP phylograms will then sometimes represent a portion only of the complete set of all MP phylograms corresponding to the inferred set of MP cladograms.

As a result, a standard MP phylogeny inference program cannot guarantee the finding of all MP solutions, even when implementing a Branch and Bound search (i.e. a search that guarantees the finding of all MP cladograms; Hendy & Penny 1982). The median-joining algorithm cannot guarantee that either (Bandelt *et al.* 1999). Although it applies the parsimony criterion locally, adding median vectors to an initial minimum spanning network to reduce the overall length of the graph, it does not explore the entire space of possible solutions. In fact, Fig. 4 shows a case where the analysis of 15 sequences by both the median-joining and maximum parsimony methods result in two different sets of most parsimonious solutions, each set containing unique solutions that the other does not. Therefore, neither method appears ideal at this time, and it might be advisable to use both methods to maximize the probability of including all MP paths in the final network. Combining the results of both approaches appears trivial: it is sufficient to add the connections present in the network graph obtained with one approach to the network graph generated by the second approach, if absent.

Conclusion

In summary, while haplotype/allele networks (defined here as a special case of phylogenetic network, see above) are often more informative than phylogenetic strict consensus trees to display intraspecific DNA sequence variation, both maximum parsimony and network-building algorithms are not guaranteed to find all most parsimonious phylograms for a set of sequences and can lead to alternative MP solutions. As already suggested in Mardulyn *et al.* (2009), considering all MP ancestral sequences of a set of MP cladograms should guarantee the inference of all most parsimonious phylograms, at least when using an algorithm exploring the space of possible trees that guarantees the finding of all MP cladograms. Developing a program capable of generating the set of all MP phylograms corresponding to a set of MP cladograms would therefore considerably improve the maximum parsimony approach and might make it superior compared with network-building algorithms, at least if we are willing to evaluate the performance of these algorithms, as I have assumed here, by measuring their ability to find all MP phylograms from a set of sequences. While the use of a criterion-based method appears desirable, the tree space to explore in the case of large data sets may become easily too large for the use of a Branch and Bound algorithm, forcing the user to turn to heuristic strategies. These still need to be compared with network-building algorithms such as the median-joining method, before deciding whether one of these approaches performs better than the other in practice.

Acknowledgements

The comments of two anonymous reviewers have significantly improved the clarity of the ideas presented here. An earlier version of the manuscript has also benefited from the careful reading of Simon Dellicour, Ceridwen Fraser, François Mayer and Maud Quinzin. This work was supported by grants from the Belgian Fonds de la Recherche Scientifique—FNRS.

References

Bandelt HJ, Dress AW (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular phylogenetics and evolution*, **1**, 242–252.

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

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *MMolecular biology and evolution*, **16**, 37–48.

Barrett CF, Freudenstein JV (2011) An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. *Molecular ecology*, **20**, 2771–2786.

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*, **21**, 255–265.

Cassens I, Van Waerebeek K, Best PB *et al.* (2003) The phylogeography of dusky dolphins (Lagenorhynchus obscurus): a critical examination of network methods and rooting procedures. *Molecular ecology*, **12**, 1781–1792.

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

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular ecology*, **9**, 1657–1659.

Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates Inc., Sunderland, Massachusetts.

Fitch WM (1997) Networks and viral evolution. *Journal of molecular evolution*, **44**, S65–S75.

Hendy MD, Penny D (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical biosciences*, **59**, 277–290.

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews. Genetics*, **4**, 275–284.

Holland B, Moulton V (2003) Consensus networks: a method for visualizing incompatibilities in collections of trees. In: *Algorithms in bioinformatics, WABI 2003* (eds Benson G and Page R), pp. 165–176. Springer-Verlag, Berlin, Germany.

Holland BR, Huber KT, Moulton V, Lockhart PJ (2004) Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular biology and evolution*, **21**, 1459–1461.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, **23**, 254–267.

Huson DH, Rupp R, Scornavacca C (2011) *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, Cambridge, UK.

Mardulyn P, Cassens I, Milinkovitch MC (2009) A comparison of methods for constructing evolutionary networks from intraspecific DNA sequences. In: *Population Genetics for Animal Conservation* (eds Bertorelle G, Bruford MD, Hauffe HC, Rizzoli A and Vernesi C), pp. 102–118. Cambridge University Press, Cambridge, UK.

Marshall DC, Hill KBR, Fontaine KM, Buckley TR, Simon C (2009) Glacial refugia in a maritime temperate climate: cicada (Kikihia subalpina) mtDNA phylogeography in New Zealand. *Molecular ecology*, **18**, 1995–2009.

Morrison DA (2005) Networks in phylogenetic analysis: new tools for population biology. *International journal for parasitology*, **35**, 567–582.

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in ecology & evolution*, **16**, 37–45.

Salzburger W, Ewing GB, Von Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular ecology*, **20**, 1952–1963.

Swofford DL (2003) *PAUP*, phylogenetic analysis using parsimony (*and other methods), v. 4b10*. Sinauer Associates, Sunderland, Massachusetts.

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

The author has a general interest in the analysis of DNA sequence variation for studying evolution. Its current research focuses mainly on comparing multilocus phylogeographic patterns among different temperate and cold-adapted herbivorous insects.

---

doi: 10.1111/j.1365-294X.2012.05622.x