Points of View

Syst. Biol. 56(5):857–862, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701633153

Haplotype Networks Can Be Misleading in the Presence of Missing Data

SIMON JOLY,[1] MARK I. STEVENS,[1,2] AND BETTINE JANSEN VAN VUUREN[3]

[1]*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand; E-mail: s.joly@massey.ac.nz (S.J.)*
[2]*School of Biological Sciences, Monash University, Clayton 3800, Victoria, Australia*
[3]*DST-NRF Centre of Excellence for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa*

Accurate haplotype networks are of critical importance for studies at the population-species interface or below the species level, in particular those that estimate parameters based on network topologies such as nested clade analyses (NCA). Although the impact of missing data in phylogenetic analyses has received much attention (reviewed in Kearney and Clark, 2003), no equivalent studies exist for haplotype network methods even though these approaches have major differences (Posada and Crandall, 2001). For example, the most commonly reported consequence of missing data in bifurcating trees is the occurrence of multiple equally likely solutions that when summarized in a consensus lead to reduced resolution in the form of polytomies (Kearney and Clark, 2003). In contrast, in haplotype networks the aim is to represent all equally most-parsimonious solutions and missing data are likely to result in greater network complexity. Moreover, because of the low levels of divergence found in intraspecific data sets, missing data have the potential to make a sequence look identical to several haplotypes, thereby reducing resolution.

Intraspecific data sets may contain ambiguous and/or missing characters for several reasons: incomplete sampling (when concatenating multiple genes or when shorter sequences are obtained for some accessions), ambiguous trace file (sequence chromatogram) because of allelic variation or poor sequencing reaction, or they might be invoked in sequence alignments containing indels longer than one base pair (insertions or deletions) to avoid giving a high weight to a single insertion or deletion (Kelchner, 2000; Simmons and Ochoterena, 2000). This latter category is expected to become more frequent with the increasing use of nuclear markers in population studies (e.g., Hare, 2001; Garrick and Sunnucks, 2006).

The behavior of three commonly used haplotype network methods was evaluated in the presence of missing data: minimum spanning networks (MSNs), statistical parsimony (SP), and full median networks (hereafter named median networks [MNs]). A complete description of algorithms for MSNs (e.g., Excoffier and Smouse, 1994; Bandelt et al., 1999), MNs (Bandelt et al., 1995) and SP (Templeton et al., 1992; Clement et al., 2002) can be found elsewhere, but a quick description of each is given as online supplemental material (http://www.systematicbiology.org).

These three methods have some important differences that need to be mentioned here. MSNs do not reconstruct any unsampled haplotypes, which implies that whenever unsampled haplotypes of degree ≥ 3 exist (the degree of a node corresponds to the number of branches connected to it), hypothetical ancestors need to be added—the Steiner problem in graph theory (Foulds et al., 1979)—and MSNs will not give optimal solutions (Fig. 1). Both SP and MNs reconstruct Steiner nodes (i.e., of degree ≥ 3) and therefore do not have the same pitfall as MSNs. MNs will always contain all most parsimonious solutions for any data set (Bandelt et al., 1995), although the downside of this is that they may also contain suboptimal solutions and that their graphical representation can be very complex. To circumvent this latter problem, modifications of MNs have been proposed (Bandelt et al., 1995, 1999; Huber et al., 2001). These modified methods will not be dealt with because they behave exactly as MNs in the situations presented here. MNs is also limited because it can only be applied to binary characters, even though a procedure to apply them to multistate characters is described elsewhere (Bandelt et al., 1995) and quasi-networks—a derivative of MNs for nucleotide data—have recently been described (Bandelt and Dür, 2007). SP aims at giving "locally" most parsimonious solutions as it gives a higher weight to short branches during network construction. Therefore, SP may exclude "real" homoplasies (if present) from shorter branches on the network, potentially resulting in a network that is not "globally" most parsimonious (S. Woolley, personal communication). This criterion could explain why inaccurate SP networks are more frequent

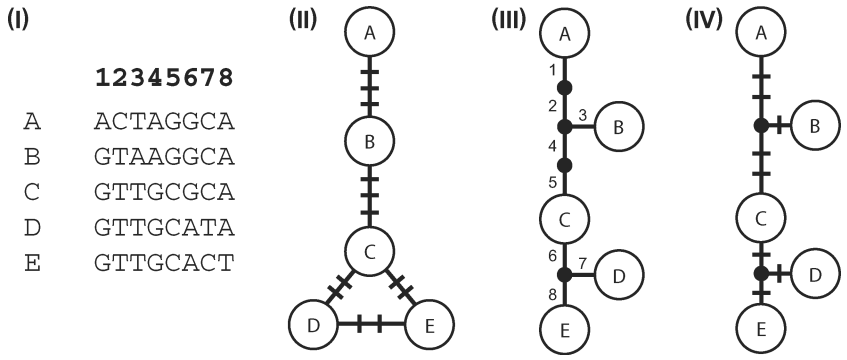

FIGURE 1. (I) Hypothetic data set used to illustrate differences among a minimum spanning network (MSN; II), a statistical parsimony network (SP; III), and a median network (MN; IV). MSNs do not infer unsampled haplotypes and fail to find the most parsimonious solution. Both MNs and SP reconstruct unsampled (or extinct) haplotypes of degree ≥3 (Steiner nodes), but SP (as obtained from the program TCS [Templeton, Crandall, and Singh]) also represents nodes of degree 2. Networks are illustrated so as to reflect their philosophies—mutations are shown as perpendicular bars along the branches for MSNs and MNs, whereas branches represent mutations for SP. Filled circles represent inferred unsampled or extinct haplotypes. Numbers above branches for SP indicate mutation events on the network; the ordering of mutation events on branches that only contain nodes of degree 2 represents one of the possible optimizations (e.g., mutations 4 and 5 can be interchanged).

when sampled haplotypes are distant from one another (Cassens et al., 2005). However, it is possible to determine if a solution is globally optimal for some data sets (Holland et al., 2004).

METHODS

Assessing the performance of network methods with missing data using simulations is not straightforward. Indeed, simulating missing data that are biologically realistic is complex because of the multiple ways that these can arise. Instead, simple hypothetical examples representing situations encountered in empirical studies were used to illustrate how these methods handle missing data. This approach makes it easier to show the impact of missing data as the effects of other confounding factors, such as homoplasy, are minimized. An alignment of six sequences that do not contain missing data (sequences A to F; Fig. 2) was analyzed to obtain expected networks. Six data sets with missing data were then constructed by individually adding a different sequence that contains missing characters (sequences G to L; Fig. 2) to the alignment without missing data (A to F). These data sets were then analyzed to see how the addition

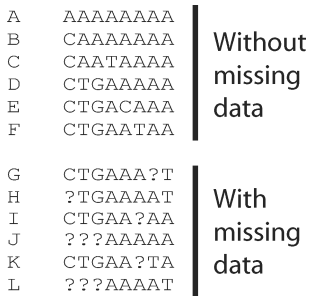

FIGURE 2. Hypothetical DNA alignment of 12 sequences used to illustrate the behavior of different algorithms for constructing haplotype networks in the presence of missing data.

of one sequence with missing data affects network reconstruction. The following situations were investigated, in which missing characters (I) are at constant sites (sequence G; Fig. 2); (II) are at variable sites and alternative optimizations do not change sequences that are at shortest distance to the one containing the missing data (sequence H; Fig. 2); (III) are at variable sites and, apart from these sites, the sequence bearing them is identical to more than one haplotype (sequences I and J; Fig. 2); and (IV) are at variable sites and alternative optimizations of missing characters change the sequences that are at shortest distance to the one containing missing characters (sequences K and L; Fig. 2). Although only one or two examples were chosen to illustrate each situation, the conclusions reached were not influenced by these choices. Moreover, the conclusions also apply to larger data sets as these situations could occur locally in a larger network.

For each of these situations, networks were constructed using MSNs, SP, and MNs. MSNs were constructed in ARLEQUIN (version 3.11; Excoffier et al., 2005) with the missing data threshold per site fixed to 25%. SP networks were inferred with the software "Templeton, Crandall and Singh" (TCS ver. 1.21; Clement et al., 2000), with the parsimony limit fixed to "2" so that all haplotypes were connected. This was necessary because all sites are variable in our example, an unrealistic assumption for intraspecific data sets. Constant sites do not alter our results and for ease of presentation additional constant sites were not added to the present examples. MNs were obtained using NETWORK (version 4.201; fluxus-engineering, 2004), with the reduced median network option but with a reduction threshold >2 to obtain a full median network. All possible re-orderings of sequences in data sets were tested to see if this influences the results.

Two criteria were used to judge the methods: (a) do missing data affect the topology of the network obtained without missing data and (b) are all most parsimonious

solutions represented for sequences with missing data that have an ambiguous position.

RESULTS

All three methods resulted in identical networks when only sequences without missing data were considered (Fig. 3). When missing data were at constant sites or when they did not change the identity of the sequences that were at the shortest distance to sequences with missing data (situations I and II), all three methods gave networks that were identical to those constructed when sequences with missing data were excluded (sequences G and H; Fig. 3). This was expected because such missing data do not affect network construction for the methods investigated.

In situations where different optimization of missing characters make a sequence identical to other (different) sequences (situation III), MNs always preserved the relationships obtained with no missing data (sequences I and J; Fig. 3). Yet, it also never showed alternative positioning of sequences with missing data. In the same situation, MSNs did not preserve the relationships obtained with no missing data (Fig. 3). In these cases where resolving missing data could make sequences identical to several different sequences, MSNs collapsed these into a single haplotype. Because of this collapsing, no alternative positioning for sequences with missing data remained.

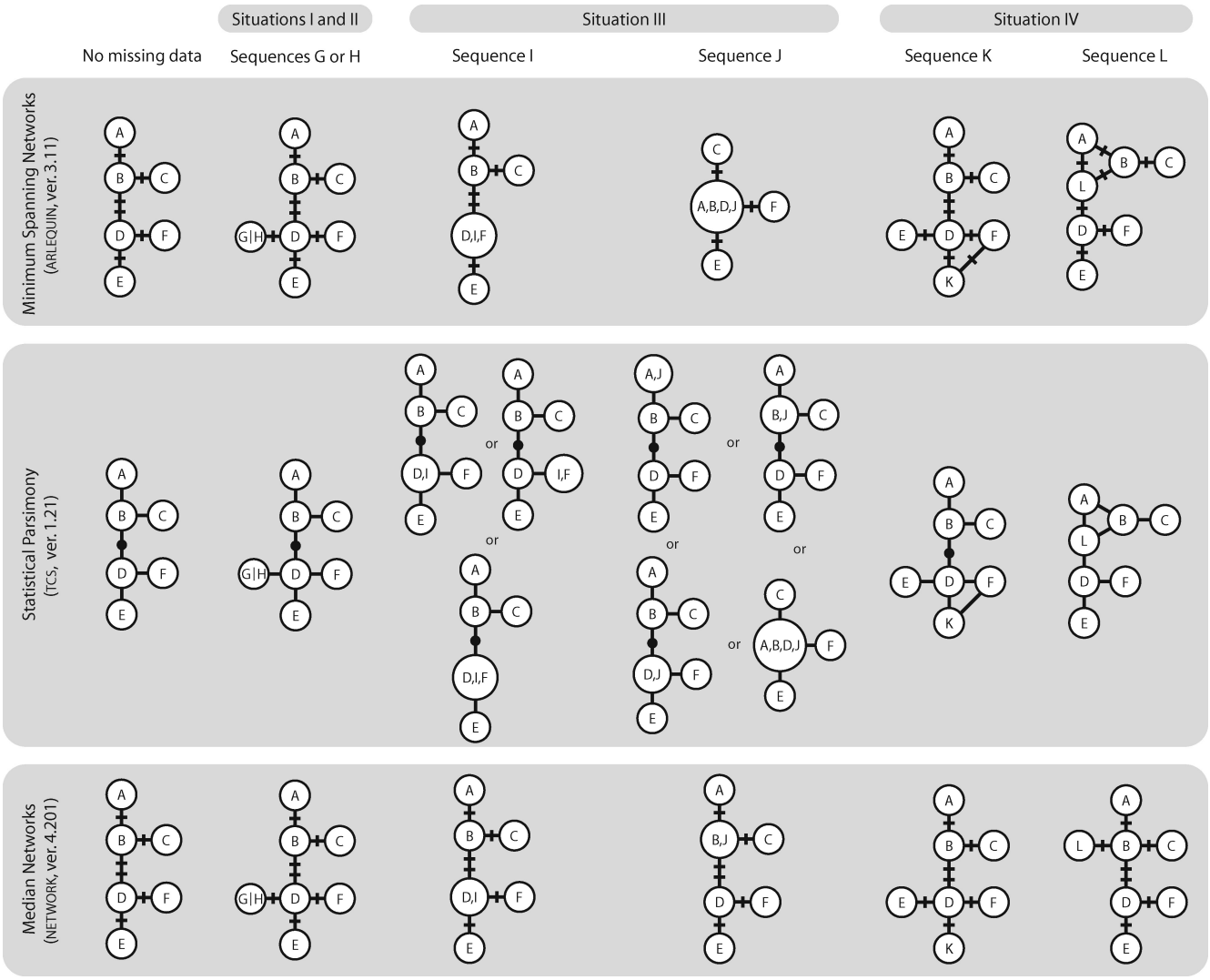

FIGURE 3. Results of network construction methods using alignments from sequences shown in Figure 2. Networks are illustrated so as to reflect their philosophies—mutations are shown as perpendicular bars along the branches for minimum spanning networks (MSNs) and median networks (MNs), whereas branches represent mutations for statistical parsimony (SP). Filled circles represent unsampled or extinct haplotypes. The situations investigated (I to IV) consist of missing characters that (I) are at constants sites; (II) are at variable sites and alternative optimizations do not change sequences that are at shortest distance to the one containing missing data; (III) are at variable sites and, apart from these sites, the sequence bearing them is identical to more than one haplotype; (IV) are at variable sites and alternative optimizations of missing characters change the sequences that are at shortest distance to the one containing missing characters.

With SP, different networks were obtained for different re-orderings of the sequences in the matrix. SP networks either showed one of the alternative positionings of the sequence with missing data or collapsed all haplotypes that had a distance of 0 with the sequence that contains missing data (Fig. 3). If the network was accurate in the former situation, it was not in the second.

In situation IV, where different optimizations of missing characters change the identity of sequences that are at shortest distance to the sequence that contain missing data (sequences K and L), only MNs preserved the relationships obtained without missing data. The results obtained with MSNs and SP when sequence K was included were correct; however, when sequence L was included results were biased. These misleading results occur when alternative most-parsimonious positions for the sequence with missing data are separated by a branch of length two on the network. MNs did not show alternative positioning of sequences with missing data, whereas MSNs and SP showed alternative positioning of sequence K and some alternatives for sequence L (Fig. 3).

The methods investigated, as currently implemented in the programs used in this study, also differed regarding the way they warn of potential problems when there are missing data in the data set. ARLEQUIN, which implements MSNs, does not give any warning when there are missing data; it only removes columns that contain more than a given percentage of missing values (the default value is 5%). NETWORK, which reconstructs MNs, reports that missing data tend to make the analysis less reliable, but it does not help to identify problematic sequences considering that not all missing data lead to problems in network construction. Finally, TCS, implementing SP, warns the user when there are ambiguities in collapsing sequences into haplotypes and lists these problematic sequences in the log file (although these lists are often incomplete), which might indicate alternative positions of sequences (e.g., situation III).

DISCUSSION

The results presented here clearly demonstrate that MSNs, SP, and MNs can all give misleading networks in the presence of missing data. In some situations, the network methods investigated resulted in biased network relationships and/or failed to indicate alternative positions for sequences. Both problems are serious because they can affect results of analyses that rely on network topology, such as population differentiation indices (i.e., using patristic distances along a network in analysis of molecular variance; Excoffier et al., 1992) or past population history inferences obtained from nested clade analyses (Templeton et al., 1995). For instance, the collapsing of different sequences into a single haplotype reduces the overall haplotype diversity and gives the illusion of shared haplotypes among populations that could bias estimates of population structure and migration rates, results obtained from nested clade analysis, and assessment of recurrent formation of polyploids or apomicts.

Results obtained for each method could be better understood if their algorithms are considered. Misleading networks occur with MSNs because ambiguous character states result in nonmetric distances by violating the triangle inequality property. Because MSNs collapse all sequences that are at distance 0 into a single haplotype, violation of the triangle inequality property by missing data can cause two distinct sequences to be collapsed into a single haplotype. For example, sequence I is at distance 0 with sequences D and F (Fig. 2), causing all three to be collapsed into a single haplotype (Fig. 3), even though the true distance between D and F is equal to 1.

MNs (as implemented in NETWORK) always preserved the relationships obtained without missing data and did not show alternative positioning for sequences because it resolves ambiguities before network construction by replacing missing data in one sequence by using the most common state found in the closest sequences (A. Roehl, personal communication), as described in Bandelt et al. (1999). In other words, all potential ambiguities are removed prior to network construction.

The results obtained with SP were highly dependent upon the order of the sequences in the data set, which is a consequence of the current implementation of SP in TCS (see also the TCS documentation; Clement et al., 2000). Consider the following alignment:

1 AT
2 AA
3 A?

TCS compares the sequences in the order 1-2, 1-3, and 2-3. In this example, the comparison 2-3 will not occur because sequence 3 has already been collapsed with 1 to give haplotype [AT]. According to the above matrix, two haplotypes will be obtained: [AT] with a frequency of 2 (sequences 1 and 3) and [AA] with a frequency of 1 (sequence 2). If the order of sequences in the matrix was instead 2-1-3, haplotypes [AT] (frequency 1; sequence 1) and [AA] (frequency 2; sequences 2-3) would be obtained. And if the first sequence of the matrix was 3, only haplotype [A?] would be obtained with a frequency of 3. Therefore, a sequence with missing data that has a distance of 0 with several distinct sequences will be grouped with the sequence that appears first in the matrix, and the other equally parsimonious alternatives will not be shown. TCS gives a warning when such ambiguities occur and identifies potential problematic sequences in the log file, but the list of ambiguous sequences is not always exhaustive. This order-dependent collapsing of sequences into haplotypes in TCS explains why different networks were obtained for SP (Fig. 3).

Although missing data were represented by question marks by convention, misleading results could also be obtained if ambiguous nucleotides were used instead. Note that if ambiguous nucleotides are used to represent a real polymorphism (e.g., because of allelic variation)

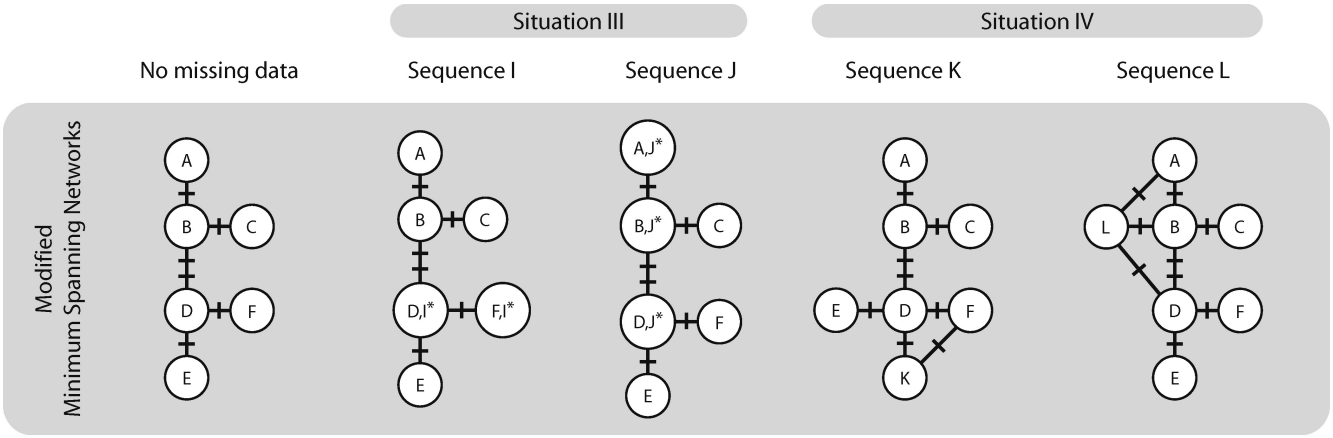

FIGURE 4. Results obtained by applying the modified MSNs described in Appendix 1 when applied on the data sets of Figure 2. An asterisk (*) beside the sequence means that its position is ambiguous; alternative positions are shown in such instances. The two situations (III and IV) consist of missing characters that (III) are at variable sites and, apart from these sites, the sequence bearing them is identical to more than one haplotype; (IV) are at variable sites and alternative optimizations of missing characters change the sequences that are at shortest distance to the one containing missing characters (see text).

rather than an unknown character state, none of the methods will reflect this information (i.e., the presence of distinct sequences) on the network.

Clearly, the results obtained here are dependent of the way each method is implemented. For example, MNs obtained from SPECTRONET (Huber et al., 2002) differ from the ones presented here because all sites with missing data are removed from the analysis. Although this removes ambiguities, it is not appealing as it reduces the amount of information present in the original data set and could collapse sequences that are different when deleted sites are considered, a property that may have important consequences for intraspecific data sets.

Suggestions

Whenever possible, potentially problematic missing data should be resolved by further experimentation. But this is not always possible and network methods need to provide ways to handle missing or ambiguous data. Network construction problems might be avoided by deleting either sequences or columns that contain most of the missing data, but as mentioned above, this solution is not desirable. Ideally, all methods should aim at fulfilling the two criteria used here, which are network accuracy and representation of alternative positions for sequences that have an ambiguous position due to missing data. However, it would also be important for methods to identify potentially problematic sequences as this would make it easier to assess how missing data may affect the results.

One simple modification could be made to MSNs that would greatly improve the results obtained with missing data. This involves constructing a network with sequences that do not have any missing data at variable sites and then subsequently adding sequences with missing data (the complete algorithm is given in Appendix 1). This modification would not collapse different sequences that do not have missing data into a single haplotype (Fig. 4), although this may occur among sequences that contain missing data.

SP could also be improved by placing all sequences with missing data at variable sites at the end of the matrix, preferably in an increasing order of missing data. As explained above, this will not collapse *different* sequences into a single haplotype, although problems may still occur if there are several sequences with missing data. Moreover, this would not show alternative positions for sequence with missing data. Again, the best strategy might be to construct a network without sequences with missing data and then adding the remaining haplotypes within the parsimony limit. Clearly, there is place for improvement with the actual algorithms regarding the handling of missing data.

In general, the present results illustrate the importance of data exploration when using network methods. Sound advice would be to analyze data sets with and without missing data to ensure that network construction is not affected and to use several network methods as they have different strengths (see also Cassens et al., 2005). Different methods will often produce different networks, either because of unsampled haplotypes, homoplasies, or missing data—a fact that warrants further attention from biologists when using network methods to obtain reliable representations of genealogies.

ACKNOWLEDGEMENTS

The authors thank S. Woolley, P. Lockhart, D. Penny, B. Holland, P. Mardulyn, M. Hedin, and an anonymous reviewer for helpful comments. S.J. was supported by a Canadian NSERC postdoctoral fellowship and MS through a NZ Foundation for Research, Science and Technology postdoctoral fellowship.

REFERENCES

Bandelt, H.-J., and A. Dür. 2007. Translating DNA data tables into quasi-median networks for parsimony analysis and error detection. Mol. Phylogenet. Evol. 42:256–271.

Bandelt, H.-J., P. G. Forster, and A. Röhl. 1999. Median-joinning networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16:37–48.
Bandelt, H.-J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743–753.
Cassens, I., P. Mardulyn, and M. C. Milinkovitch. 2005. Evaluating intraspecific "network" construction methods using simulated sequence data: Do existing algorithms outperform the global maximum parsimony approach? Syst. Biol. 54:363–372.
Clement, M., D. Posada, and K. A. Crandall. 2000. TCS: A computer program to estimate gene genealogies. Mol. Ecol. 9:1657–1659.
Clement, M., Q. Snell, P. Walker, D. Posada, and K. A. Crandall. Year. TCS: Estimating gene genealogies. 2002. *in* First IEEE International Workshop on High Performance Computational Biology (HiCOMB), Fort Lauderdale, Florida. Available at http://www.hicomb.org/HiCOMB2002/.
Excoffier, L., L. G. Laval, and S. Schneider. 2005. Arlequin, version 3: An integrated software package for population genetics data analysis. Evol. Bioinform. Online 1:47–50.
Excoffier, L., and P. E. Smouse. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. Genetics 136:343–359.
Excoffier, L., P. Smouse, and J. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131:479–491.
fluxus-engineering. 2004. Network, version 4.201. www.fluxus-engineering.com/sharenet.htm.
Foulds, L. R., M. D. Hendy, and D. Penny. 1979. A graph theoretic approach to the development of minimal phylogenetic trees. J. Mol. Evol. 13:127–149.
Garrick, R. C., and P. Sunnucks. 2006. Development and application of three-tiered nuclear genetic markers for basal hexapods using single-stranded conformation polymorphism couple with targeted DNA sequencing. BMC Genetics 7:11.
Hare, M. P. 2001. Prospects for nuclear gene phylogeography. Trends Ecol. Evol. 16:700–706.
Holland, B. R., K. T. Huber, D. Penny, and V. Moulton. 2004. The Min-Max squeeze: Guaranteeing a minimal tree for population data. Mol. Biol. Evol. 22:235–242.
Huber, K. T., M. Langton, D. Penny, V. Moulton, and M. Hendy. 2002. Spectronet: A package for computing spectra and median networks. Appl. Bioinformatics 1:159–161.
Huber, K. T., V. Moulton, P. J. Lockhart, and A. Dress. 2001. Pruned median networks: A technique for reducing the complexity of median networks. Mol. Phylogenet. Evol. 19:302–310.
Kearney, M., and J. M. Clark. 2003. Problems due to missing data in phylogenetic analyses including fossils: A critical review. J. Vertebr. Paleontol. 23:253–274.
Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann. Mo. Bot. Gard. 87:482–498.
Posada, D., and K. A. Crandall. 2001. Intraspecific gene genealogies: Trees grafting into networks. Trends Ecol. Evol. 16:37–45.
Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49:369–381.
Templeton, A. R., K. A. Crandall, and C. F. Singh. 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.
Templeton, A. R., E. Routman, and C. A. Phillips. 1995. Separating population structure from population history: A cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. Genetics 140:767–782.

First submitted 10 May 2007; reviews returned 26 June 2007; final acceptance 10 July 2007
Associate Editor: Marshal Hedin
Editor-in-Chief: Jack Sullivan

APPENDIX 1 AN ALGORITHM FOR CONSTRUCTING MINIMUM SPANNING NETWORKS IN THE PRESENCE OF MISSING DATA

This modified minimum spanning network (MSN) algorithm first constructs a network only from sequences that do not have missing data and then adds sequences with missing data: (1) calculate the distance between any two sequences and arrange these in increasing order; (2) put all distance comparisons that include sequences with missing data at variable sites aside to first consider only comparisons between sequences that do not have missing data. Let δ_i be the smallest distance observed in the sample; (3) connect all haplotypes of *different* sub-networks that are at distance δ_i; (4) repeat step (3) with second shortest distance (δ_{i+1}) and so on until a network of all haplotypes is formed; (5) then consider comparisons with sequences that have missing data; (6) starting with the smallest distance (δ_j), connect all haplotypes of *different* sub-networks that are at distance δ_j. When a sub-network (or haplotype) can be connected at more than one place on another sub-network at a given distance, connect it to all of them; (7) repeat step (6) with the next shortest distance (δ_{j+1}) until all sequences with missing data are connected to the network.

Syst. Biol. 56(5):862–870, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701636412

The Drowning of New Zealand and the Problem of *Agathis*

MICHAEL KNAPP,[1,5] RAGINI MUDALIAR,[2] DAVID HAVELL,[3] STEVEN J. WAGSTAFF,[4] AND PETER J. LOCKHART[1]

[1]*Institute of Molecular BioSciences, Allan Wilson Centre, Massey University; E-mail: michael_knapp@eva.mpg.de (M.K.)*
[2]*School of Biological Sciences, University of the South Pacific, Suva, Fiji*
[3]*Department of Conservation, Auckland, New Zealand*
[4]*Manaaki Whenua Landcare Research, PO Box 69, Lincoln 8152, New Zealand*
[5]*Current Address: Department of Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany*

Eighty million years ago (Ma) the landmass that was to become New Zealand broke away from the Gondwanan supercontinent. During the subsequent Oligocene period (26 to 38 Ma), there was a significant reduction in the landmass of New Zealand (Cooper and Cooper, 1995, and references therein). However, whether or not New Zealand was completely submerged is a matter of controversy and recent debate