Syst. Biol. 54(1):56–65, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590906046

Biogeographic Interpretation of Splits Graphs: Least Squares Optimization of Branch Lengths

RICHARD C. WINKWORTH,[1] DAVID BRYANT,[2] PETER J. LOCKHART,[3] DAVID HAVELL,[4] AND VINCENT MOULTON[5]

[1]*Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA; E-mail: Richard.Winkworth@yale.edu*
[2]*Department of Mathematics, McGill University, Canada; E-mail: bryant@mcb.mcgill.ca*
[3]*Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand; E-mail: p.j.lockhart@massey.ac.nz*
[4]*Universal College of Learning, Palmerston North, New Zealand; E-mail: D.Havell@ucol.ac.nz*
[5]*The Linnaeus Centre for Bioinformatics, University of Uppsala, Box 598, 751 24, Uppsala, Sweden; E-mail: Vincent.Moulton@lcb.uu.se*

Abstract.—Although most often used to represent phylogenetic uncertainty, network methods are also potentially useful for describing the phylogenetic complexity expected to characterize recent species radiations. One network method with particular advantages in this context is split decomposition. However, in its standard implementation this approach is limited by a conservative criterion for branch length estimation. Here we extend the utility of split decomposition by introducing a least squares optimization technique for correcting branch lengths that may be underestimated by the standard implementation. This optimization of branch lengths is generally expected to improve divergence time estimates calculated from splits graphs. We illustrate the effect of least squares optimization on such estimates using the Australasian *Myosotis* and the Hawaiian silversword alliance as examples. We also discuss the biogeographic interpretation and limitations of splits graphs. [Biogeography; hybridization; least squares; recent species radiation; reticulation; split decomposition.]

Phylogenetic networks are important tools for studying complex patterns in molecular sequence data. Amongst other applications, they have been used to study intraspecific DNA sequence variation (e.g., Bandelt et al., 1995, 2000), viral and bacterial evolution (e.g., Worobey et al., 2002; Kotetishvili et al., 2002), and plant species diversity (e.g., Huber et al., 2001; Lockhart et al., 2001). In such cases phylogenetic networks have advantages over tip-labeled bifurcating evolutionary models because (1) ancestral sequences are often present in the population of extant sequences, and (2) uncertainty in phylogenetic reconstruction can be easily visualized (Bandelt et al., 1995; Holland and Moulton, 2003). Indeed these properties make networks useful for studying any biological system in which the evolutionary process is expected to be complex and nonbifurcating.

Plant species radiations have the potential to provide important insights into the process of plant speciation. Evolutionary studies on species radiations generally model diversification as a bifurcating process. However, given that hybridization, introgression, and polyploidy may play important roles in the rapid evolution of species diversity in such groups (e.g., Lockhart et al., 2001), it seems likely that this view is too simplistic. Furthermore, the evolution of the multiallelic and multilocus nuclear markers that are often used to investigate closely related species are also likely to be complex and not well described by a bifurcating model (e.g., Sota and Vogler, 2003). Here we illustrate the phylogenetic complexity typical of plant species radiations and the usefulness of networks in such cases with an example from the New Zealand alpine flora. Specifically, we focus on three closely related yet morphologically, ecologically, and geographically distinct species of *Ranunculus* ("buttercups") endemic to the mountains of southern New Zealand (Fisher, 1965; Webb et al., 1988). *Ranunculus sericophyllus* is widely distributed along the western southern alps of New Zealand's South Island (Fig. 1a), occurring on wet stony ground at the snowline fringe (1500 to 2150 m). In contrast, *R. pachyrrhizus* occurs on the drier southeastern mountains of Otago (South Island; Fig. 1a), along tarn edges or associated with melt-water at the snowline (1200 to 2150 m). The most localized of these species, *R. viridis*, is restricted to rocky ledges, clefts, and hollows in the subalpine zone of Mt. Allen in the Tin Range of Stewart Island (700 m; Fig. 1a). Visual inspection of an alignment for nuclear ribosomal ITS (nrITS) sequences from 20 accessions indicates character state differences that distinguish the three species and within *R. sericophyllus* differentiate between three geographic regions—the central Southern Alps, northern Fiordland, and southern Fiordland (Fig. 1b; Lockhart et al., 2001; Lockhart, unpublished). However, even if the heteroplasmic sites are removed from the sequence alignment prior to phylogenetic analysis these distinctions are not well represented by standard parsimony or maximum likelihood trees (Fig. 1c and d). In this example the patterns of incompatibility among nucleotide sites are too complex to be modeled by a bifurcating tree. In contrast, phylogenetic networks can represent these patterns of sequence variation. For example, Figure 1e shows a splits graph constructed using the standard implementation of split decomposition in SplitsTree4.0 (Huson and Bryant, 2004) and p-distances calculated from the same data as used for Figure 1c and d.

Posada and Crandall (2001) provide a review of network methods and their advantages for analyzing intraspecific gene genealogies. Many of the benefits of network approaches in this context also extend to the study of recent species radiations. For example, phylogenetic relationships within species radiations are not expected to be hierarchical but instead to be characterized by low genetic divergence, the persistence of ancestral sequence types, as well as multifurcate and reticulate patterns of evolution. Network methods provide an effective means of representing such situations because they are capable of displaying more of the phylogenetic information

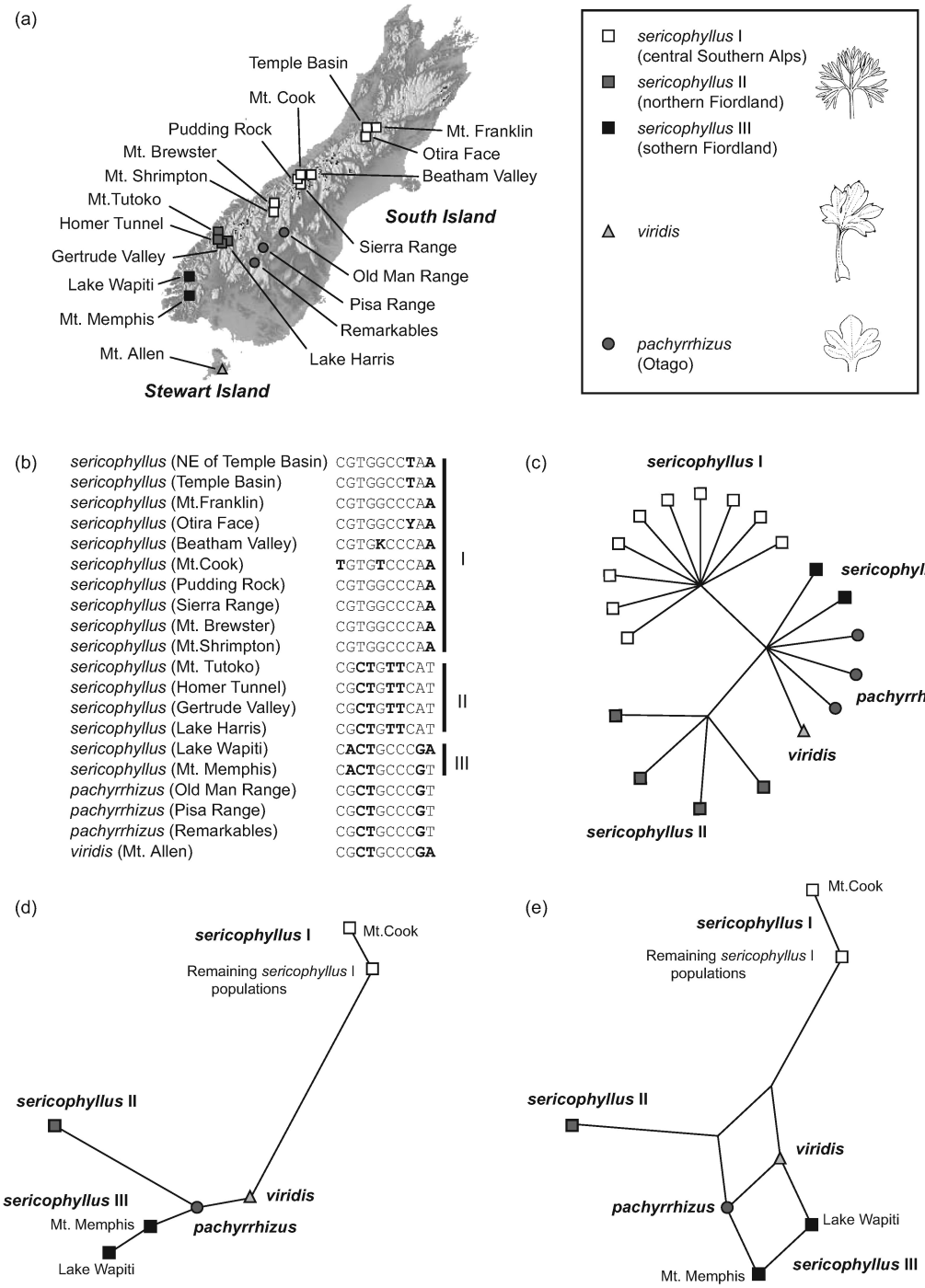

FIGURE 1. (a) Map of the New Zealand's South Island and Stewart Island showing collection localities for three endemic species of *Ranunculus—R. pachyrhizus*, *R. sericophyllus*, and *R. viridus*. (b) Alignment of varied nucleotide positions in nrITS sequences for 20 accessions of these three species. Locations correspond to those in (a); roman numerals denote three phylogeographically distinct groups of *R. sericophyllus*. (c) Strict consensus of 19 most parsimonious trees using PAUP* 4.0b10 (Swofford, 2002). (d) One of two optimal ML trees from a heuristic search of the complete nrITS sequences using PAUP* 4.0b10; the second tree differed in the placement of the *sericophyllus* II clade, which was instead attached at the *R. viridus* node. The substitution model (TVMef+I; Lset, Base = equal, Nst = 6, Rmat = [0.0000 2571931.0000 4618441.0000 0.0000 2571931.0000], Rates = equal, Pinvar = 0.9800) was selected by AIC in Modeltest Version 3.06c (Posada and Crandall, 1998). (e) Splits graph constructed with *p*-distances using SplitsTree4.0 (beta 06; Huson and Bryant, 2004).

contained in a data matrix; in particular, they can visualize potentially competing signals. Various network methods are available; examples include median networks (Bandelt et al., 1995, 2000), median-joining networks (Bandelt et al., 1999), reticulograms (Legendre and Makarenkov, 2002; Makarenkov and Legendre, 2004), split decomposition (Bandelt and Dress, 1992), and statistical parsimony (Templeton et al., 1992).

Although network methods in general are useful for studying recent species radiations, split decomposition has two specific advantages in this context. Firstly, the method is canonical; that is, there is a single unique solution represented by the splits graph. This property is particularly useful because potentially competing solutions can be visualized simultaneously. In contrast, the reticulogram approach (e.g., Makarenkov and Legendre, 2004) is not canonical. The resulting reticulogram can vary depending on the initial input tree, making it difficult to evaluate alternative solution. The second advantage of split decomposition is that only the strongest signals of incompatibility are represented in the splits graph. This limits the visual complexity of the graph and facilitates biological interpretation even when levels of incompatibility are high. Like split decomposition, the related median network method (e.g., Bandelt et al., 1995) is also canonical; however, median graphs can be visually highly complex because all patterns of incompatibility are represented graphically. Despite these advantages, in its standard implementation the utility of split decomposition may be limited by a conservative criterion for branch length selection. Although this conservative approach ensures only well-supported edges are represented in the splits graph, it has the disadvantage that branch lengths in the graph are likely to systematically underestimate the distances calculated directly from the data set. The problem is particularly acute when numerous quartets are associated with a given edge (e.g., the quartets AB|CD and AB|CE are both associated with the branch AB|—) because the chance of an overly conservative estimate occurring in the split system is greater. This bias may have important implications for formulating and testing biogeographic hypotheses; especially in the context of divergence time estimation because the ages calculated from standard splits graphs will likely be underestimates.

In this article we describe a procedure for the least squares optimization of branch lengths in a phylogenetic network. Optimized graphs are expected to provide more accurate estimates of branch length and therefore better represent relationships between the sequences. Differences between standard and optimized splits graph networks are illustrated using two examples. We also discuss the interpretation of splits graphs in the context of plant biogeography and species radiations.

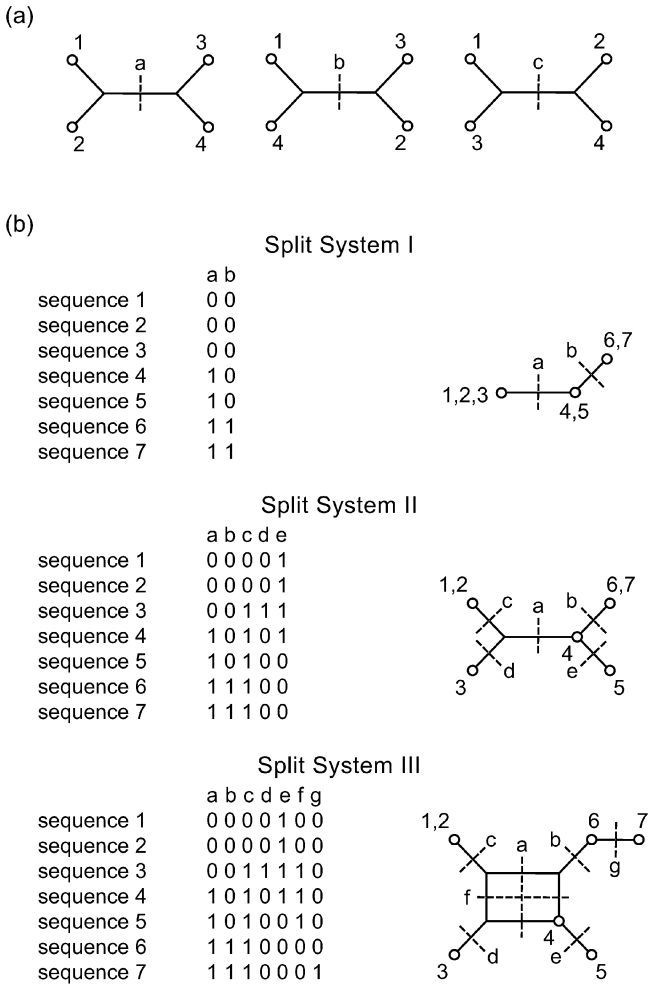

FIGURE 2. (a) The three possible unrooted topologies for the taxa 1, 2, 3, and 4. The splits a, b, and c uniquely define the quartets. (b) A simple example of splits graph construction. The splits described in the split systems are represented in the corresponding graphs. Modified from Lockhart et al. (2001).

MATERIALS AND METHODS

An Introduction to Split Decomposition

Split decomposition is a transformation-based approach for visualizing evolutionary data (Bandelt and Dress, 1992; Huson, 1998). The method decomposes the data to a "sum of weakly compatible splits," which is then visualized as a splits graph. In the context of DNA sequences, the method first considers the three possible unrooted tree topologies for all quartets of sequences in the data set (Fig. 2a). For each of these trees support for the *internal split* is evaluated using, in the standard implementation, a distance calculation. The so-called *isolation index* is given by

$$0.5 - (d_{14} + d_{23} - d_{12} + d_{34})$$

where the d_{ij} are path lengths between pairs of taxa. For each quartet the two trees with the highest isolation index are retained and included in the *split system*—simply the collection of weakly compatible splits that will be used to assemble the splits graph. The split system also contains splits that describe the *external edges*. Support for these is evaluated by considering all possible triplets of

sequences. If the taxa are labeled i, j, k then the isolation index for the split leading to i is given by

$$0.5 - (d_{ij} + d_{ik} - d_{jk})$$

Obviously several quartets (or triplets, in the case of external edges) may describe the same split, and so it is often necessary to chose between possible values of the isolation index. In the standard implementation of split decomposition, the smallest value of the isolation index is used to represent a given edge. Once the splits have been identified and their isolation index values determined, a graph is constructed using an algorithm that progressively separates the sequences from one another (Fig. 2b). If the data have evolved under a divergent process and do not contain incompatible patterns of nucleotide substitutions, then the resulting splits graph will be treelike. However, if such incompatibilities are present then these will be represented as reticulations in the reconstructed network. A goodness of fit metric, called the *split decomposition fit statistic*, is also calculated and provides an indication of how well the graph represents the original distances. The statistic is the sum of pairwise distances represented in the graph divided by the sum of those observed in the data. A fit of 100% indicates that the graph fully represents the distances in the data set (e.g., Fig. 1e); lower fit values, suggesting poor correspondence between the graph and the data, are expected when levels of nucleotide incompatibility are high.

Least Squares Optimization

In standard splits graphs the length of a given edge corresponds to the smallest value of the isolation index for that split. This conservative selection criterion tends to negatively bias branch length estimates and may result in graphs that are a poor fit to the distances calculated from the original data. Least squares approaches are statistically well justified and widely applicable to data fitting and optimization problems (Felsenstein, 1984, 2003). Indeed the intuition behind applying least squares to phylogenetic trees and networks is the same: we wish to find branch lengths for the graph that most closely fit the distances (either corrected or uncorrected) estimated from the data. For a phylogenetic tree the distance between two taxa is simply the sum of the branch lengths along the path that connects them. This is also true of networks, except that in this case the shortest distance between two taxa may be measured along several possible paths (see Fig. 3). In this section we describe a least squares fitting procedure for the optimization of edge lengths on a splits graph. We also describe goodness of fit measures that can be used to evaluate such optimized graphs.

Under the ordinary least squares (OLS) framework we measure the fit between the estimated distances, p_{ij}, in a graph and the observed distances (or corrected distances), d_{ij}, by

$$\text{SS(OLS)} = \sum_{ij} (p_{ij} - \Sigma d_{ij})^2 \tag{1}$$

(Cavalli-Sforza and Edwards, 1967). We may also increase the influence of certain terms in this summation by applying the weighted least squares (WLS) measure

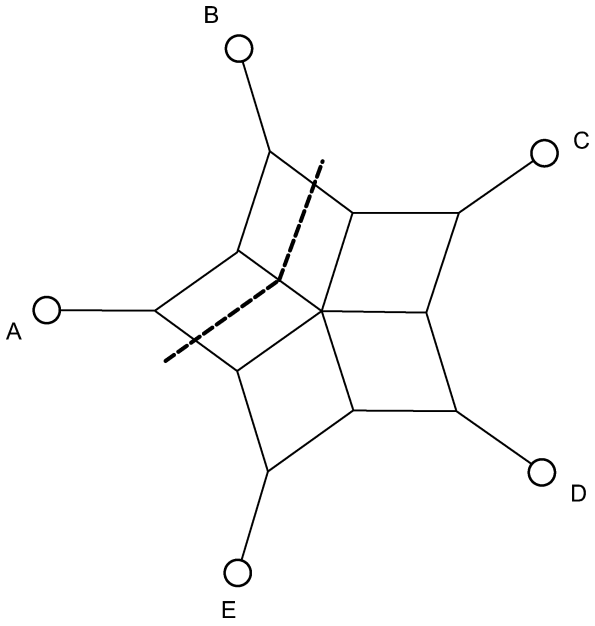

FIGURE 3. A splits graph representing 10 splits for 5 terminals. Each split is represented by a set of parallel branches. For example the split {A, B} | {C, D, E} corresponds to the parallel branches crossed by the dotted line. The length for each split, S, is proportional to the isolation index, α^s (Bandelt and Dress, 1992).

$$\text{SS(WLS)} = \sum_{ij} w_{ij}(p_{ij} - \Sigma d_{ij})^2 \tag{2}$$

where the w_{ij} are positive weight values typically derived from variances. Fitch and Margoliash (1967) suggest the weightings $w_{ij} = 1/d_{ij}$ or $w_{ij} = 1/d_{ij}^2$, as these approximate the reciprocal of the variance of d_{ij}.

With both trees and networks we can use linear algebra to determine the branch lengths that minimize SS(OLS) or SS(WLS). We use the same notation as Rzhetsky and Nei (1993) and others. Let N be the number of taxa. The observed and estimated distances are both stored in the $N(N-1)/2$ dimensional vectors

$$\mathbf{d} = (d_{12}, d_{13}, \ldots, d_{(N-1),N})' \tag{3}$$

and

$$\mathbf{p} = (p_{12}, p_{13}, \ldots, p_{(N-1),N})' \tag{4}$$

respectively (we use $'$ to denote transpose). Let m be the number of splits in the tree or network. For a tree each branch corresponds to a different split, so m also equals the number of branches in the tree. In contrast, splits graphs generally consist of more branches than splits because, in a network, a split may correspond to a collection of parallel branches (Fig. 3); however, note that all branches corresponding to a single split have the same length. For both trees and networks we use a vector

$$\mathbf{b} = (b_1, b_2, \ldots, b_m)' \tag{5}$$

to store the branch lengths.

When estimating least squares branch lengths on a tree we express the shape of the tree in terms of a topological matrix **A**. The matrix **A** has $N(N-1)/2$ rows (one for each pair of taxa i, j) and m columns (one for each branch). We put a 1 in column k and the row for pair i, j if the path between i and j passes over branch k. Otherwise, we place a 0 at this matrix position. Hence, $\mathbf{A}_{(ij)k}$ is 1 when i and j are on different sides of the split corresponding to branch k. Because a network also represents a set of splits, this characterization immediately extends to splits graphs. In this context, the topological matrix **A** for a splits graph has $N(N-1)/2$ rows (one for each pair of taxa i, j) and m columns (one for each split).

Subsequently, the distances (p_{ij}) between taxa in the tree or network are determined from the branch lengths, **b**, by the formula $\mathbf{p} = \mathbf{Ab}$. The branch lengths minimizing SS(OLS) are found by solving the linear equation

$$\mathbf{A}'\mathbf{Ab} = \mathbf{A}'\mathbf{d} \tag{6}$$

a formula dating to Gauss; or to Cavalli-Sforza and Edwards (1967) in the context of phylogenetic trees. For weighted least squares we construct the $N(N-1)/2$ by $N(N-1)/2$ dimensional matrix, **W**, with the values w_{ij} on the diagonal and zeros everywhere else. The branch lengths minimizing SS(WLS) are then given by solving

$$\mathbf{A}'\mathbf{WAb} = \mathbf{A}'\mathbf{Wd} \tag{7}$$

a formula applied to phylogenetics by Fitch and Margoliash (1967) and Farris (1972).

There are many methods for solving linear equations such as (6) or (7); a comprehensive survey of these has been made by Golub and van Loan (1996). The computer program SplitsTree4.0 implements the method of Cholesky decomposition (Golub and van Loan 1996; c.f. page 8) with algorithms for solving linear equations specifically designed for the (positive definite) matrices $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}'\mathbf{WA}$. However, as data sets increase in size and complexity more sophisticated algorithms may be required to evaluate the network.

One property of the standard implementation of split decomposition is that the sum of distances in a splits graph will always be equal to or less than the sum of the distances in the original distance matrix; this provides the basis for the split decomposition fit statistic. However, this measure may be invalid when branch lengths are optimized using the least squares procedure, because the sum of distances in the graph may, in some cases, exceed the sum of the distances in the original matrix. This problem can be remedied by using an alternative statistic,

$$fit_{\text{SDiff}} = 1 - \sum_{ij} \left| p_{ij} - d_{ij} \right| \Big/ \sum_{ij} d_{ij} \tag{8}$$

which we will call the *sum of differences goodness of fit*. This statistic is equivalent to the split decomposition fit statistic when all observed distances are greater than those estimated in the graph. However, because it is defined generally, it remains valid even when the estimated distances exceed the observed values for some pairs of sequences. Again a fit of 100% indicates that the observed and estimated distances coincide exactly.

In the context of least squares a more widely used measure of fit is that introduced by Tanuka and Huba (1985),

$$fit_{LS} = 1 - \sum_{ij} (p_{ij} - d_{ij})^2 \Big/ \sum_{ij} d_{ij}^2 \tag{9}$$

which we will call the *least squares goodness of fit*. As for both previous measures, the maximum value of statistic (i.e., 100%) occurs only when the observed and inferred distances coincide exactly.

Examples

We investigated differences between standard and least squares optimized splits graphs, and more specifically the effect of branch length optimization on divergence time estimation, using nrITS sequence data from *Myosotis* (Boraginaceae; Winkworth et al., 2002) and the Hawaiian silversword alliance (Asteraceae; Baldwin and Robichaux, 1995). For each data set multiple sequence alignments were performed using ClustalX, with visual inspection. Prior to phylogenetic analyses all ambiguous and gapped positions were excluded from data matrices (available as supplementary materials from the Systematic Biology website). For *Myosotis* we used a HKY85+I correction for distance calculations; the model and parameters were those used by Winkworth et al. (2002). Baldwin and Sanderson (1998) used a HKY85+G substitution model for maximum likelihood analyses of nrITS sequences from the Hawaiian silverswords and relatives. Our data sets differ somewhat from those used in that study and so we tested for a best-fit model using Modeltest Version 3.06c (Posada and Crandall, 1998); a GTR+I+G model was selected for a data set including outgroups, and an HKY model best-fit the ingroup-only data set. However, these tests rely on a bifurcating representation of the data (i.e., Modeltest uses a neighbor-joining tree) and our preliminary analyses suggested that these data were not treelike. Rather than attempt to choose between models in this case we instead consider uncorrected distances. Standard and least squares optimized splits graphs, as well as the corresponding goodness of fit measures, were calculated using SplitsTree4.0 (beta 06; Huson and Bryant, 2004).

So that our age estimates were directly comparable to those from the earlier studies we used different approaches to divergence time estimation for each data set. For *Myosotis* we estimated the age of the austral group using the total (HKY85+I) distance between the most recent common ancestor of Northern and Southern Hemisphere lineages and *Myosotis exarrhena*, which was identified as the most divergent austral taxon in the maximum likelihood analysis of Winkworth et al. (2002). The corresponding edges in the splits graphs are labeled and thickened in Figure 4. To calculate absolute time we

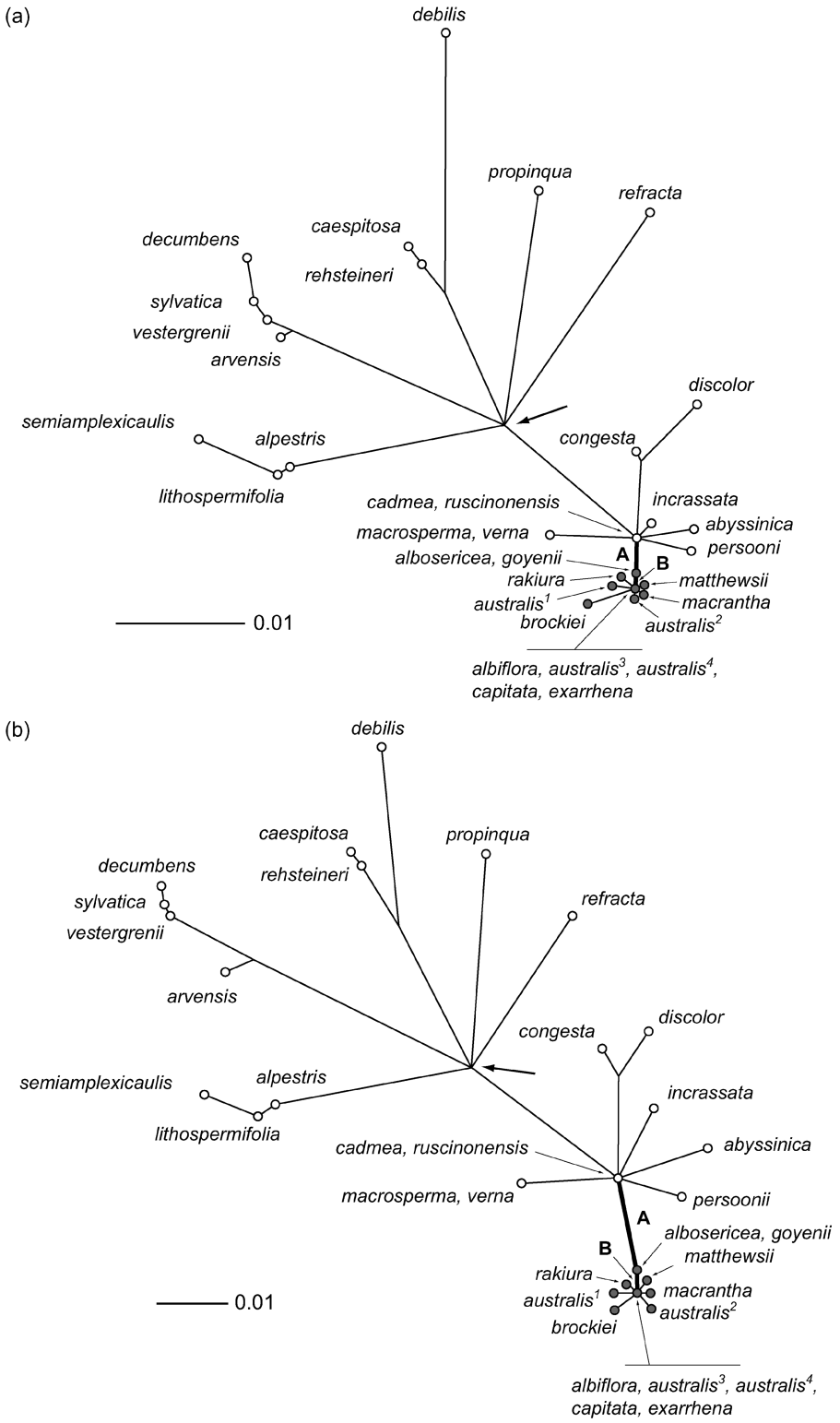

FIGURE 4. Splits graphs for nrITS sequences from 34 *Myosotis* taxa, constructed using HKY85+I distances estimated in PAUP*4.0b10 (Swofford, 2002). GenBank accession numbers for sequences are given in Winkworth et al. (2002). (a) Branch lengths estimated using the standard implementation of split decomposition (SplitsTree4.0 beta 06; Huson and Bryant, 2004). Shaded circles indicate austral taxa. *Myosotis australis* is represented by four accessions—[1]Mt. Kozciuscko (Australia), [2]New Guinea, [3]Tasmania, [4]*australis* "yellow" (New Zealand). Sum of differences goodness of fit = 46.1%, least squares goodness of fit = 71.75%. Thickened arrow indicates placement of outgroups when included in analyses. (b) Branch lengths optimized using the least-squares function as implemented in SplitsTree4.0 (beta 06; Huson and Bryant, 2004). Australasian taxa, *M. australis* accessions, and outgroup root position denoted as for (a). Sum of differences goodness of fit = 93.31%, least squares goodness of fit = 99.37%. In both graphs edges used in divergence time estimation (see text) are thickened and labeled.

used the most conservative evolutionary rate suggested by Winkworth et al. (2002); a fossil pollen–based calibration of 1.10×10^{-9} substitutions/site/year. To test the robustness of our estimates we constructed confidence intervals using nonparametric bootstrapping. For 1000 replicates we recorded the lengths of edges A and B (assigning length zero if the corresponding split did not appear in a specific replicate). Total length estimates (i.e., the sum of A and B) were ranked and the total width of the 95% confidence intervals given by age calibrating replicates 25 and 975.

In contrast Baldwin and Sanderson (1998) calculated an average age for the radiation of the Hawaiian silversword. Specifically, they estimate the age of this radiation using the average divergence from the most recent common ancestor of the Hawaiian lineage. In order to root our networks, we conducted preliminary analyses that included the outgroup taxa *Madia madioides*, *Madia bolanderi*, *Raillardiopsis muirii*, and *Raillardiopsis scabrida*. We then measured the total distance between the node corresponding to the root and each terminal (i.e., the sum of edge lengths along the, or one of the, shortest paths between these points) in graphs that contained only Hawaiian taxa. The average divergence within the Hawaiian silversword alliance was calculated and age calibrated using an evolutionary rate from Richardson et al. (2001; 3.00×10^{-9} substitutions/site/year). Confidence intervals for our estimates were again constructed using nonparametric bootstrapping.

Results and Discussion

Utility of Splits Graphs for the Study of Recent Species Radiations

Molecular phylogenetic analyses suggest that for many plant groups contemporary species diversity has been strongly influenced by late Tertiary and Quaternary events. Particularly striking are dramatic morphological and ecological radiations that appear to be correlated with Quaternary climatic fluctuations (e.g., Comes and Kadereit, 1998; Kadereit et al., 2004) or recent colonization of insular environments (e.g., Baldwin, 1992; Böhle et al., 1996). Often these radiations are characterized by hybridization and introgression, and in some cases also by polyploidization (Schaal et al., 1998). Indeed, in specific cases the evidence indicates that hybridization has been an important process in the adaptive radiation of plant lineages (Rieseberg et al., 2003). If shown to be a general phenomenon, then plant species radiations are likely to be characterized by complex patterns of phylogenetic relationship.

It is well recognized that reticulate evolution (e.g., hybridization and polyploidy) can confound phylogeny reconstruction because different marker loci may have different histories. For plants, processes such as introgression (e.g., Rieseberg and Wendel, 1993), "chloroplast capture" (e.g., Soltis et al., 1991; Whittlemore and Schaal, 1991), and genome reorganization (e.g., Song et al., 1995; Rieseberg et al., 2003) are potential outcomes of reticulate evolution. However, hybridization and polyploidy may also influence molecular evolution at specific loci. For example, the presence of heteroplasmic nucleotide positions in nrITS sequences from putative hybrids has been interpreted as reflecting the failure of concerted evolution to homogenize differentiated parental ITS repeats following hybridization (Sang et al., 1995a; Sota and Vogler, 2003). Disruption of the correction mechanism may also allow incomplete gene conversion or crossing over to recombine diverged parental sequences. As a result, hybrids display novel sequence types containing character states derived from both the maternal and paternal lineages (Buckler et al., 1997; Aguilar et al., 1999; Sota and Vogler, 2003). Clearly these character incompatibilities cannot be mapped onto a single bifurcating tree; but such situations can be visualized using a phylogenetic network. Potential examples are illustrated in Figures 1e and 5; in both cases the observed reticulations may have resulted from introgressive hybridization and recombination between nrITS sequences.

It is important, when drawing inferences from splits graphs, to bear in mind that the internal nodes of a splits graph are not necessarily equivalent to those in a bifurcating tree. When nucleotide site patterns are pairwise compatible, the internal nodes of a splits graph will correspond to ancestral sequences (as they generally do in a bifurcating tree). However, if incompatibility is high then internal nodes may not represent ancestors. Instead they are simply vertices required for construction of the splits graph. Consequently, reticulations in splits graphs should generally be considered indicators of phylogenetic complexity rather than diagnostic of specific evolutionary events. Although this conservative interpretation may appear biologically unsatisfying, keep in mind that bifurcating evolutionary models are unlikely to provide an unambiguous reconstruction of phylogenetic relationships or ancestral states using such data. For example, in Figure 1 both the parsimony and maximum likelihood searches recovered multiple trees. Presenting these topologies in the form of a strict consensus tree (Fig. 1c) results in a loss of phylogenetic information, whereas reporting a single tree does not fully represent the underlying complexity (Fig. 1d). In contrast, the splits graph (Fig. 1e) displays the complex relationships between the sequences, and, despite our conservative interpretation of the reticulations, provides a framework for evolutionary inference.

Branch Length Estimation in Splits Graphs

Although a standard splits graph may represent complex evolutionary relationships more fully than a bifurcating tree, branch length estimates in such graphs may differ substantially from those calculated using a global reconstruction method such as least squares or maximum likelihood. Specifically, the conservative criterion for selecting branch lengths in a standard splits graph is likely to systematically underestimate distances in the original data set. This is particularly problematic in the context of divergence time estimation because ages calculated from such graphs would also be underestimates.

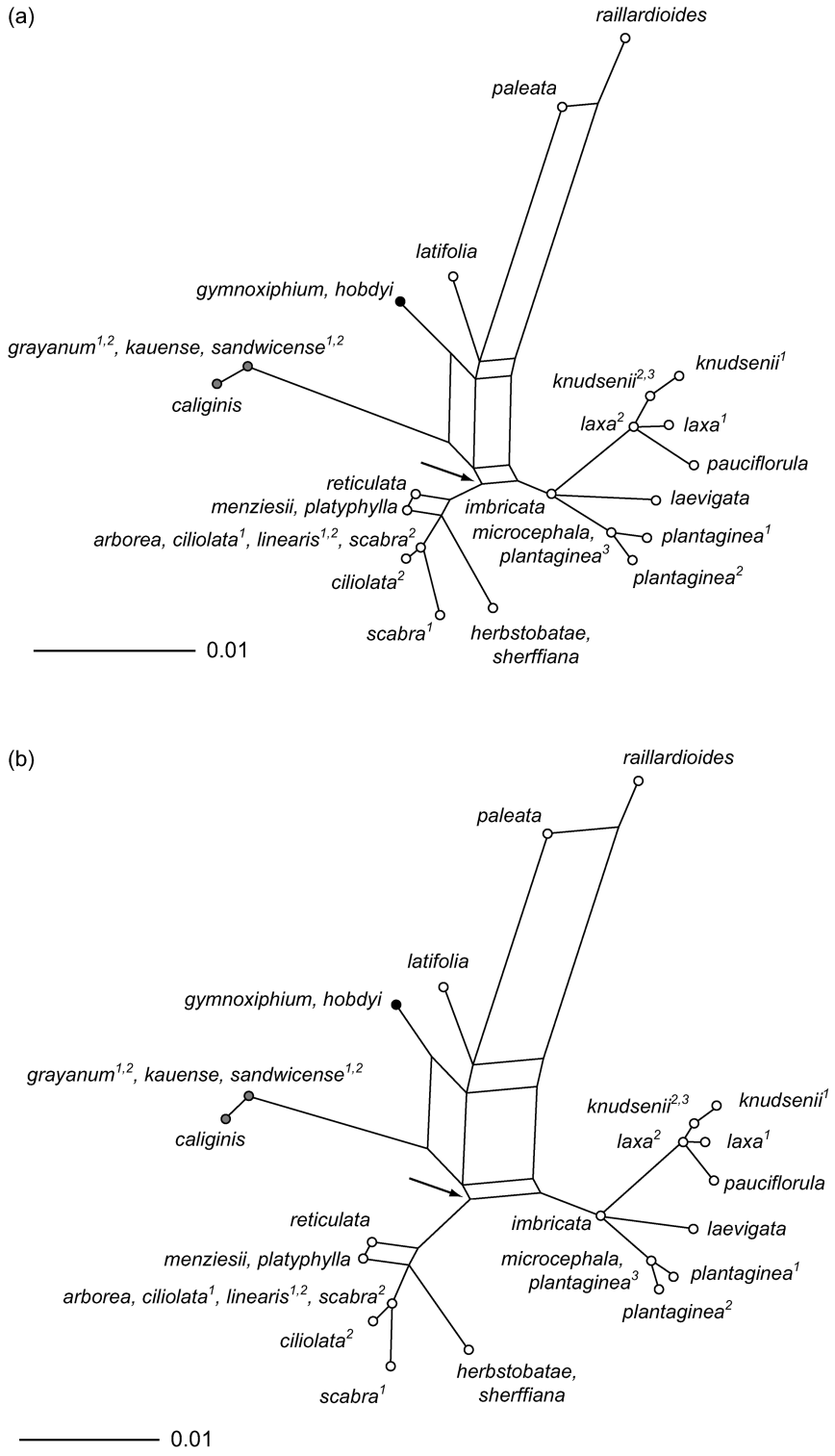

FIGURE 5. Splits graphs constructed from *p*-distances from a nrITS data set for the Hawaiian silversword alliance. GenBank accession numbers for sequences are given in Baldwin and Sanderson (1998). (a) Branch lengths estimated using the standard implementation of split decomposition (SplitsTree4.0 beta 06; Huson and Bryant, 2004). The three silversword genera are indicated—grey, *Argyroxiphium*; white, *Dubautia*; black, *Wilkesia*. Eight taxa are represented by multiple accessions: *A. grayanum*—[1]East Maui, [2] West Maui; *A. sandwicense*—[1]subsp. *sandwicense*,[2]subsp. *macrocephalum*; *D. ciliolata*—[1]subsp. *glutinosa*, [2]subsp. *ciliolata*; *D. knudsenii*—[1]subsp. *knudsenii*, [2]subsp. *filiformis*, [3]subsp. *nagatae*; *D. laxa*—[1]subsp. *hirsuta*, [2]subsp. *laxa*; *D. linearis*—[1]subsp. *linearis*,[2]subsp. *hillebrandii*; *D. plantaginea*—[1]subsp. *plantaginea*, [2]Kaua'i, [3]subsp. *humilis*; *D. scabra*—[1]subsp. *leiophylla*, [2]subsp. *scabra*. Sum of differences goodness of fit = 77.33%, least squares goodness of fit = 94.19%. Thickened arrow indicates placement of outgroups when included in analyses. (b) Branch lengths optimized using the least squares function as implemented in SplitsTree3.2. Silversword genera, multiple accessions, and outgroup root placement denoted as for (a). Sum of differences goodness of fit = 96.16%, least squares goodness of fit = 99.79%.

In contrast, least squares optimization is expected to produce splits graphs that better represent the relationships between sequences because edge lengths are not systematically biased. We illustrate these differences using Australasian *Myosotis* (Winkworth et al., 2002) and the Hawaiian silversword alliance (Baldwin and Robichaux, 1995).

Recent molecular phylogenetic analyses of nrITS and chloroplast *matK* sequences have improved our understanding of relationships within *Myosotis* (the "forget-me-nots"). This study suggested that the morphologically diverse Australasian taxa are a monophyletic group that arrived recently by long-distance dispersal from the northern hemisphere (Winkworth et al., 2002). Using maximum likelihood estimation (with an HKY+I model of evolution) on the optimal bifurcating tree for the nrITS sequences, these authors conservatively estimated that the austral lineage had diverged from its northern hemisphere relatives approximately 14.7 million years ago (Mya). We constructed standard (Fig. 4a) and least squares optimized (Fig. 4b) splits graphs using the *Myosotis* nrITS sequences of Winkworth et al. (2002). The standard splits graph for these data is treelike; however, the relatively low value of the split decomposition fit statistic (i.e., 46.1%) suggests that some nucleotide incompatibilities in the original data set are not represented by the graph. Least squares optimization provided a substantial improvement in the goodness of fit. Specifically, the sum of differences goodness of fit increased from 46.1% to 93.31%, whereas the least squares goodness of fit increased from 71.75% to 99.37%. A visual comparison also indicates pronounced changes in the relative lengths of several branches following least squares optimization of the splits graph. For example, edge A, which subtends the austral radiation, is more than four times longer in the optimized splits graph (branch length = 0.01315) than in the standard graph (branch length = 0.00267). This difference has a substantial impact on the inferred age of the Australasian lineage. Age estimates from the standard splits graph suggest that the divergence of northern and southern hemisphere *Myosotis* occurred 4.4 Mya (95% bootstrap confidence interval 0–8.6 Mya). In contrast, the least squares optimized graph suggests this event is 14.8 Myr, old (95% bootstrap confidence interval 6.7–26.1 Myr), an age much closer to that suggested by the maximum likelihood analysis.

The Hawaiian silversword alliance—consisting of *Argyroxiphium*, *Dubautia*, and *Wilkesia*—is perhaps the best-known and well-studied botanical example of adaptive radiation on an oceanic island archipelago. Evolutionary studies on this morphologically and ecologically diverse group have considered cytogenetic, isozymic, and DNA variation. Using the nrITS sequences reported by Baldwin and Robichaux (1995), we constructed standard and optimized splits graphs. These graphs indicate that the evolution of the nrITS has not been strictly treelike in this group (Fig. 5a and b). As in the previous example, least squares optimization improved the fit of branch lengths in the network to those observed from the data. The standard graph had a sum of differences goodness of fit of 77.33% compared to 96.16% for the optimized graph; the least squares goodness of fit statistic improves from 94.19% to 99.79%. Visual comparison of the splits graphs indicates changes in relative branch length that correspond to differences in inferred divergence times. Assuming an evolutionary rate of 3.00×10^{-9} substitutions/site/year (from Richardson et al., 2001), edge length estimates from the standard splits graph (Fig. 5a) suggest an average age of 3.9 Myr, (95% bootstrap confidence interval 2.9–5.9 Myr) for the most recent common ancestor of the Hawaiian silverswords. In contrast, the least squares optimized graph (Fig. 5b) suggests that diversification began 5.3 Mya (95% bootstrap confidence interval 3.7–7.3 Mya). This latter estimate is closer to that of Baldwin and Sanderson (1998), who suggest the most recent common ancestor of the silversword group was 5.2 ± 0.8 Myr old based on a penalized likelihood analysis of nrITS sequences (using a HKY85+G substitution model).

These examples indicate that least squares optimization (1) improves the fit between input distances and branch lengths in the reconstructed graph, and (2) leads to age estimates that more closely match those calculated from standard maximum likelihood approaches. Although we expect least squares optimization to generally improve edge length estimation, the procedure will be ineffective in two situations. Specifically, (1) if the fit of the distances is already 100%, or (2) if, under the standard split decomposition implementation, support for the internal split is zero (i.e., the split has length 0). This latter problem may arise if the process of substitution is not uniform across the underlying phylogeny or when some sequences are very divergent, and consequently the errors in distance estimation are large (e.g., Adachi and Hasegawa, 1996; Phillippe and Douzery, 1994; Ranwez and Gascuel, 2001).

CONCLUSIONS

The complex evolutionary processes that often characterize plant species radiations are not likely to be well represented by bifurcating tree models. In contrast, phylogenetic networks provide a powerful tool for exploring the extent and distribution of incompatibilities because they are capable of graphically representing the competing signals in a data set. Here we extend the utility of split decomposition by implementing a least-squares optimization procedure for estimating branch length, which we show improves the fit between the input distances and the resulting splits graph. In general we expect this approach will lead to improved estimates for divergence times and therefore more realistic inferences about historical biogeography and species radiations.

ACKNOWLEDGEMENTS

We thank the New Zealand Marsden Fund, The Alexander von Humboldt Foundation, and Massey University for their financial support of our studies on the New Zealand flora. The authors also thank Rod Page, Sebastian Bocker, David Posada, and Chris Simon for helpful comments on an earlier version of the manuscript.

REFERENCES

Adachi, J., and M. Hasegawa. 1996. Instability of quartet analyses of molecular sequence data by the maximum likelihood method: The cetacea/artiodactyla relationships. Mol. Phylogenet. Evol. 16:72–76.

Aguilar, J. F., J. A. Rossello, and G. Nieto Feliner. 1999. Nuclear ribosomal DNA (nrDNA) concerted evolution in natural and artifical hybrids of *Armeria* (Plumbaginaceae). Mol. Ecol. 8:1341–1346.

Baldwin, B. G. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the Compositae. Mol. Phylogenet. Evol. 1:3–16.

Baldwin, B. G., and R. H. Robichaux. 1995. Historical biogeography and ecology of the Hawaiian silversword alliance. Pages 259–287 *in* Hawaiian biogeography: Evolution on a hot spot archipelago (W. L. Wagner and V. A. Funk, eds.). Smithsonian Institution Press, Washington, D.C.

Baldwin, B. G., and M. J. Sanderson. 1998. Age and diversification of the Hawaiian silversword alliance (Compositae). Proc. Natl. Acad. Sci. USA 95:9402–9406.

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

Bandelt, H.-J., P. Forster, and A. Röhl. 1999. Median-joining 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.

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

Bohle, U.-R., H. H. Hilger, and W. F. Martin. 1996. Island colonisation and evolution of the insular woody habit in *Echium* L. (Boraginaceae). Proc. Natl. Acad. Sci. USA 92: 11740–11745.

Buckler, E. S., A. Ippolito, and T. P. Holtsford. 1997. The evolution of ribosomal DNA: Divergent paralogues and phylogenetic implications. Genetics 145:821–832.

Cavalli-Sforza, L. L., and A. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 21:550–570.

Comes, H. P., and J. W. Kadereit. 1998. The effect of Quaternary climatic changes on plant distribution and evolution. Trends Pl. Sci. 3:431–438.

Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. 106:645–668.

Felsenstein, J. 1984. Distance methods for inferring phylogenies: A justification. Evolution. 38:16–24.

Felsenstein, J. 2003. Inferring phylogenies. Sinauer, Sunderland, Massachusetts.

Fisher, F. J. F. 1965. The alpine Ranunculi of New Zealand. Botany Division, Department of Scientific and Industrial Research, Wellington.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. Science 155:279–284.

Golub, G., and C. Van Loan. 1996. Matrix Computations, 4th Edition. John Hopkins University Press, Baltimore, Maryland.

Holland, B., and V. Moulton. 2003. Consensus networks: A method for visualising incompatibilities in collections of trees. Proceedings of the Workshop on Algorithms in Bioinformatics (WABI), September 15–20, 2003 in Budapest, Hungary.

Huber, K., 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.

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

Huson, D. H., and D. Bryant. 2004. SplitsTree4.0 beta 06. Distributed by the authors (http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome_en.htm).

Kadereit, J. W., E. M. Griebeler, and H. P. Comes. 2004. Quaternary diversification in European alpine plants: pattern and process. Phil. Trans. R. Soc. Lond. B 359:265–274.

Kotetishvili, M., O. C. Stine, A. Kreger, J. G. Morris, and A. Sulakvelidze. 2002. Multilocus sequence typing for characterization of clinical and environmental salmonella strains. J. Clin. Microbiol. 40:1626–1635.

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

Lockhart, P. J., P. A. McLenachan, D. Havell, G. Glenny, D. Huson, and U. Jensen. 2001. Phylogeny, dispersal and radiation of New Zealand alpine buttercups: Molecular evidence under split decomposition. Ann. Mo. Bot. Gard. 88:458–477.

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

Philippe, H., and E. Douzery. 1994. The pitfalls of molecular phylogeny based on four species, as illustrated by the cetaceae/artiodactyla relationship. J. Mamm. Evol. 2:133–152.

Posada, D., and K. A. Crandall. 1998. Modeltest: Testing the model of DNA substitution. Bioinformatics 14:817–818.

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

Ranwez, V., and O. Gascuel. 2001. Quartet-based phylogenetic inference: Improvements and limits. Mol. Biol. Evol. 18:1103–1116.

Richardson, J. E., R. T. Pennington, T. D. Pennington, and P. M. Hollingsworth. 2001. Rapid diversification of a species-rich genus of neotropical rain forest trees. Science 293:2242–2245.

Rieseberg, L. H., O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, J. L. Durphy, A. E. Schwarzbach, L. A. Donovan, and C. Lexer. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301:1211–1216.

Rieseberg, L. H., and J. F. Wendel. 1993. Introgression and its consequences in plants. Pages 70–109 *in* Hybrid zones and the evolutionary process (R. G. Harrison, ed). Oxford University Press, Oxford.

Rzhetsky, A., and M. Nei. 1993. Theoretical foundation of the minimum evolution method of phylogenetic inference. Mol. Biol. Evol. 10:1073–1095.

Sang, T., D. J. Crawford, and T. F. Stuessy. 1995. Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: Implications for biogeography and concerted evolution. Proc. Natl. Acad. Sci. USA 92:6813–6817.

Schaal, B. A., D. A. Hayworth, K. M. Olsen, J. T. Rauscher, and W. A. Smith. 1998. Phylogeographic studies in plants: Problems and prospects. Mol. Ecol. 7: 465–474.

Soltis, D. E., P. S. Soltis, T. G. Collier, and M. L. Edgerton. 1991. Chloroplast DNA variation within and among genera of the Heuchera group (Saxifragaceae): Evidence for chloroplast transfer and paraphyly. Am. J. Bot. 78:1091–1112.

Song, K., P. Lu, K. Tang, and T. C. Osborn. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc. Natl. Acad. Sci. USA 92:7719–7723.

Sota, T., and A. P. Vogler. 2003. Reconstructing species phylogeny of carabid beetles *Ohomopterus* using multiple nuclear DNA sequences: Heterogeneous information content and the performance of simultaneous analyses. Mol. Phylogenet. Evol. 26:139–154.

Swofford, D. L. 2002. PAUP*. Phylogenetic Analysis using parsimony (* and other methods), version 4.10b, Sinauer, Sunderland, Massachusetts.

Tanaka, J. S., and G. J. Huba. 1985. A fit index for covariance structure models under arbitrary GLS estimation. Brit. J. Math. Stat. Psychol. 38:197–201.

Templeton, A. R., K. A. Crandall, and C. F. Sing. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetic 132: 619-633.

Webb, C. J., W. R. Sykes, and P. J. Garnock-Jones. 1988. Flora of New Zealand, Vol. 4 CRI. Landcare Research, New Zealand.

Whittlemore, A. T., and B. A. Schaal. 1991. Interspecific gene flow in oaks. Proc. Natl. Acad. Sci. USA 88:2540–2544.

Winkworth, R. C., A. W. Robertson, J. Grau, and P. J. Lockhart. 2002. Biogeography of the cosmopolitan genus *Myosotis (Boraginaceae)*. Mol. Phylogenet. Evol. 24:180–193.

Worobey, M., A. Rambaut, O. G. Pybus, D. L. Robertson, M. L. Gibbs, J. S. Armstrong, and A. J. Gibbs. 2002. Questioning the evidence for genetic recombination in the 1918 "Spanish Flu" virus. Science 296:211.

First submitted 7 August 2003; reviews returned 4 January 2004; final acceptance 18 August 2004
Associate Editor: Rod Page