Using Data-Display Networks for Exploratory Data Analysis in Phylogenetic Studies

David Morrison

Research Article

Using Data-Display Networks for Exploratory Data Analysis in Phylogenetic Studies

David A. Morrison

Section for Parasitology (SWEPAR),

Swedish University of Agricultural Sciences,

751 89 Uppsala,

Sweden

E-mail: David.Morrison@bvf.slu.se

Tel: +46 18 674164

Key words: phylogenetic networks, EDA, splits networks

Running head: Exploratory data analysis in phylogenetics

Abstract

Exploratory data analysis (EDA) is a frequently under-valued part of data analysis in biology. It involves evaluating the characteristics of the data *before* proceeding to the definitive analysis in relation to the scientific question at hand. For phylogenetic analyses, a useful tool for EDA is a data-display network. This type of network is designed to display any character (or tree) conflict in a dataset, without prior assumptions about the causes of those conflicts. The conflicts might be caused by (a) methodological issues in data collection or analysis, (b) homoplasy, or (c) horizontal gene flow of some sort. Here, I explore 13 published datasets using splits networks, as examples of using data-display networks for EDA. In each case, I performed an original EDA on the data provided, to highlight the aspects of the resulting network that will be important for an interpretation of the phylogeny. In each case, there is at least one important point (possibly missed by the original authors) that might affect the phylogenetic analysis. I conclude that EDA should play a greater role in phylogenetic analyses than it has done.

Introduction

There has been considerable recent interest in the use of networks rather than trees as the basis for phylogenetic analysis. The intention is to replace the Darwinian model of a bifurcating tree by a "reticulating tree", with the reticulations representing evolutionary processes other than lineal descent with modification. Such process involve gene flow of some sort, including: hybridization, introgression, recombination, horizontal / lateral gene transfer, genome fusion, ancestral polymorphism / deep coalescence / incomplete lineage-sorting, and gene duplication–loss.

Unfortunately, this field is rather poorly developed at the moment (Morrison 2010; Nakhleh 2010). Networks that try explicitly to represent evolutionary history (called evolutionary networks) all have serious restrictions on the types of patterns they can analyze, and on the allowed complexity of those patterns. As noted by Huson et al. (2009): "there are many promising directions to follow and rudimentary software implementations, [but] there is no tool currently available that biologists could easily and routinely use on real data."

What we have, instead, is a wide array of methods for displaying data conflict in phylogenetic datasets (called data-display networks). That is, compatible data patterns are displayed as a tree, while incompatibilities in the data are displayed as reticulations in the tree. The data may be either raw character data (e.g. sequences, AFLP, microsatellites, SNPs) or they may be characters summarized as a set of trees (e.g. gene trees). The importance of the distinction between these two types of network has been re-iterated in the literature (Nakhleh et al. 2003; Bryant and Moulton 2004; Morrison 2005, 2010; Huson and Bryant 2006; Reeves and Richards 2007; Ayling and Brown 2008), along with the possible role of data-display networks in exploratory data analysis.

The main issue here is that *any* data conflict can create reticulations in a data-display network, irrespective of its source. In addition to the gene-flow processes listed above, incompatibilities can arise from (Morrison 2010): (i) homoplasy due to analogous rather than homologous characters

(e.g. parallelism, convergence, reversal); or (ii) methodological issues in data collection (e.g. taxon sampling, character sampling, outgroups) or data analysis (e.g. model mis-specification, choice of optimality criterion). These patterns may confound the search for (and display of) the gene-flow processes being analyzed in an evolutionary network.

The detection of data conflicts, and the extent to which data conflicts will affect the data analysis, then becomes an important first step in a phylogenetic analysis. Mathematically, this is exploratory data analysis (EDA) or descriptive data analysis. EDA is often an undervalued tool in biological studies (Ellison 2001; Behrens and Yu 2003), although see Grant and Kluge (2003) for a contrary point of view. Instead, biologists frequently proceed directly to (statistical) hypothesis testing (inferential or confirmatory data analysis) without considering the nature of their data or the suitability of the test for those data. However, it is not prudent to rely on mathematical tests without a detailed exploration of the data first (Bandelt 2005). The use of networks as a tool for EDA has rarely been illustrated in the scientific literature.

My purpose in this paper is therefore to illustrate the valuable role of data-display networks in EDA as part of a phylogenetic analysis. Data-display networks may reveal reticulation patterns that are unsuspected in the data, and which may have an important bearing on subsequent analyses and their interpretation. I do not consider the ways in which the reticulation patterns should be dealt with in the subsequent parts of the phylogenetic analysis: my aim is simply to highlight methods by which such patterns can be detected.

I do this using empirical examples of phylogenetic analyses from the literature. In each case, I have performed an original EDA on the data provided, and I highlight the aspects of the resulting network that will be important for an interpretation of the phylogeny. In each case, there is at least one important point missed by the original authors, which might affect the phylogenetic analysis. For convenience, I have restricted myself to data-display networks based on splits graphs (Huson and Bryant 2006). I start with a consideration of EDA and splits graphs, and then proceed to the

role of EDA in studying technical problems, reticulation processes, unresolved phylogenies, and possible limitations of analyses.

Methods

Exploratory Data Analysis

EDA traditionally involves both graphical displays of the data and numerical summaries of the data (Tukey 1977). The objective is not just to summarize the patterns in the data but to visualize them in a way that is meaningful in a phylogenetic context. These ways should highlight potential problems with any phylogenetic analysis of the dataset, and preferably do so in a way that allows rapid interactive assessments of the data.

Here, I am concerned solely with graphical displays, in this case a reticulated tree. Mathematically, a tree is an acyclic, leaf-labeled and connected line graph. For the purposes here, the graph can additionally have undirected cycles (reticulations). The cycles are undirected because there is no obligation to interpret the graph as an evolutionary diagram, and so the entire graph is undirected (or unrooted, in biological parlance). I will refer to a labeled graph with possible reticulations as a network. The network contains nodes linked by edges. The nodes do not necessarily represent ancestors (as they would in a rooted tree), and the edges do not necessarily represent biological character transformations (from ancestor to descendant). At least some of the nodes may simply be serving a heuristic role in portraying data conflict, and the edges simply represent apparent differences in data between the nodes (due to any cause). I will stress this by using “edge” rather than “branch”.

The reason for emphasizing networks in EDA of phylogenetic data is that if sequences are analyzed by fitting them to a tree then the output will be a tree, by definition, regardless of whether

the data are tree-like or not. However, if these data are fitted to a network, then the output will be somewhere between the two extremes of a binary tree and a completely reticulating network. If the sites in the data are compatible with one another, then the output will be a tree. Alternatively, if there is incompatibility among some (or all) of the sites, then the output will be a form of network. For this reason alone it would be very wise to survey phylogenetic data using network methods before attempting to infer phylogenetic trees.

The range of available techniques for constructing data-display networks has been surveyed recently by a number of authors (Posada and Crandall 2001; Morrison 2005, 2010; Vriesendorp and Bakker 2005; Huson and Bryant 2006; Makarenkov et al. 2006; Gemeinholzer 2008), and so I will not enumerate them here. Data-display networks, of whatever type, are fast and relatively easy to calculate, which makes them ideal as a tool for EDA.

There is no set protocol for EDA, unlike hypothesis testing where there is a strict need for an *a priori* hypothesis and usually a formal mathematical procedure. In many ways, EDA is a "fishing expedition", where a number of lures are tried in order to detect the presence of something of particular interest (e.g. a fish rather than an old boot). There may be *a priori* ideas about what interesting things there are to be found (e.g. evidence of gene-flow processes), but there is also the intention to catch whatever is available (i.e. other patterns of data conflict), and keep it if it is over the pre-determined size limit (e.g. large patterns that may affect the subsequent interpretation of the data). The metaphor is thus a good one.

For EDA in phylogenetic analysis, then, the idea is to try a few different network-construction methods, and see what they produce. To this end, convenience is as valid a criterion as any for the choice of methods, pending more detailed information about the relative merits of the available algorithms. In what follows, I have simply used a range of splits-network methods available in the program SplitsTree either v. 2.4 (Huson 1998) or 4.3 (Huson and Bryant 2006), and I display the network(s) that seem to me to be most relevant to the particular points that I wish to make. This use

of splits networks follows the work on EDA in phylogenetics by Holland et al. (2004, 2005) and Wägele and Mayer (2007).

Splits Networks

There are many algorithms now available for generating splits networks (Huson and Bryant 2006; Morrison 2010), including: (a) median networks (Bandelt 1994; Bandelt et al. 2000), split decomposition (Bandelt and Dress 199), parsimony splits (Bandelt and Dress 1993) and neighbor-net (Bryant and Moulton 2002, 2004) for character or distance data; (b) consensus networks (Holland and Moulton 2003; Holland et al. 2005, 2006) and super-networks (Huson et al. 2004, 2006) for data represented as multiple trees. Also, recombination networks (Huson and Kloepper 2005) and hybridization networks (Huson et al. 2005; Huson and Kloepper 2007) can be derived from splits networks as meta-analyses for evolutionary networks. Here, I concentrate on group (a). For some detailed examples of EDA using group (b) see Holland et al. (2004, 2005).

Note that my possibly arbitrary choice from among these method is unlikely to unduly influence the interpretation of the EDA. For example, if there is a large amount of conflicting character information in a dataset, then the median network will have a large complex set of cycles, while the split decomposition and parsimony splits networks will be unresolved, and the neighbor-net will be somewhere in between. These are all valid ways of representing the conflict, but in this case the neighbor-net will be the one that is clearest to interpret.

As an introductory example of the use of a splits network, Fig. 1 shows a parsimony splits analysis of Table 1, which contains data for 16 phenotypic characteristics that might be useful for reconstructing the evolutionary history of 12 extant vertebrate groups. The network edges have a simple 1:1 relationship to the character data, as they also would for a maximum-parsimony tree (you can easily confirm this for yourself). There are 13 characters used in the network analysis

(invariant and non-binary characters are ignored), and 12 of these form a perfect series of nested sets, which can be represented as a tree. However, character 15 (homeothermy) is incompatible with characters 12–14, forming a single undirected cycle.

In a splits network, an undirected cycle consists of two sets of parallel edges. Each set of edges represents a single bipartition (or split) of the samples. For example, the edges in Fig. 1 labeled with character 15 separate the partition {Birds, Mammals} from the remaining taxa; and the edges labeled with characters 12–14 separate the partition {Birds, Crocodiles} from the remaining taxa. Note that in this particular example all of the nodes represent extant taxa. This is unusual, as usually there will be unlabeled nodes formed by the junctions of the sets of parallel edges — these should not interpreted as unobserved "ancestors".

Ideally, there would be a simple interpretation of a splits network, with tree-like structure representing unconflicting character patterns, cycles representing conflicting patterns, and unresolved structure representing lack of character information. Unfortunately, this simple interpretation is often confounded by the fact that, if there are too many conflicting patterns to be displayed, then this will also be shown as an unresolved structure.

Note that we should not interpret Fig. 1 as an evolutionary diagram. It might be tempting to do so, perhaps rooting the network on the Lamprey-Shark edge (in a manner analogous to using an outgroup to root a phylogenetic tree). The Birds might then be interpreted as "hybrids" between Mammals and Crocodiles! However, all the network graph is doing in this case is highlighting the character conflicts, which might be better interpreted here as homoplasy (i.e. parallel origin of homeothermy in birds and mammals).

Results and Discussion

Methodological Problems in Data Analysis

There is a seemingly endless list of known or potential methodological problems associated with phylogenetic analysis, presumably arising from the fact that we are trying to reconstruct historically unique events given only contemporary data. This may be the hardest sort of mathematical analysis that a biologist can try. Here, I illustrate the use of data-display networks to investigate three well-known issues.

It is standard procedure these days (other than for parsimony analysis) to decide on which substitution model to use based on some formal procedure, in order to identify the "best fitting" model, which is the one that will then be used. The rationale for this approach is that the best-fitting model is the one that is least likely to introduce methodological problems.

This point is illustrated in Fig. 2. The left-hand network shows use of the WAG+G amino-acid substitution matrix, which is the best-fitting one available for this example dataset, and the right network shows the MtMAM+G matrix, which is the worst-fitting one, both as assessed by the ProtTest program v. 1.3 (Abascal et al. 2005). While much of the structure of the two splits networks is basically the same, the relative sizes of the splits change as the fit of the model worsens, some new minor splits are added and one major one is deleted ({Plasmodium, Arabidopsis} versus the rest). Thus, there is a clear difference in the assessed complexity of the character conflicts between the two amino-acid models, with the best-fitting model being substantially "cleaner". This shows that choosing an inappropriate model can definitely make a phylogenetic analysis harder to interpret than it needs to be.

It is worth mentioning that model mis-specification has long been recognized as a problem in tree-building analyses, and there is no a priori reason to expect network analyses to be any more

robust. Evolutionary analyses of molecular data require both a graph model to be specified and a substitution model. Networks generalize the first type of model but do not necessarily influence the second type. For example, compositional heterogeneity (Jermiin 2004) might create exactly the same sorts of problems for network analyses as for trees.

Another perennial problem in a phylogenetic analysis is the stability of the junction between the outgroup and the ingroup. Outgroups can be either too distant from or too close to the ingroup to be effective determinants of the true root. Divergent outgroups will be affected by stochastic variation (so that their location is almost random), while close outgroups may not be reciprocally monophyletic with the ingroup (so that there is no single root location).

This point is illustrated in Fig. 3. The structure of the splits network from the ingroup alone (top figure) is reflected in the network containing both the ingroup and outgroup (bottom figure), but the intricacy of the network cycles is much greater when the outgroup is included. Indeed, there is a 3-dimensional cycle, as well as several extra 2-dimensional ones. This indicates that the outgroup has complex relationships to the ingroup, and the root of any phylogenetic tree is likely to be unstable. It is perhaps unsurprising that the groups are not reciprocally monophyletic here, given that the data have been sampled at the population level (8 worms from each of 8 farms).

One of the most notorious manifestations of the opposite rooting problem is long-branch attraction (LBA) caused by distant outgroups (Bergsten 2005). The use of data-display networks as part of an EDA in search of LBA has previously been explored by Kennedy et al. (2005) and Wägele and Mayer (2007).

Fig. 4 shows an example based on trying to find the root of the angiosperms. The neighbor-joining tree (Fig. 4a) shows a classic case of possible LBA, with the long edges leading to the outgroup (the edge labeled b) and the grasses (edge a) being separated by a very short edge (e) from the edges leading to the remaining monocotyledons (c) and the dicotyledons (d). The implication here is that the root might be on the wrong edge, with the distant outgroup being attracted to the

longest ingroup edge. If the tree is rooted on edge d, instead, then the monocots and dicots would both be monophyletic. Alternatively, the most popular choice in the literature is that the root should be somewhere in the group joined to edge c. The use of a median network (Fig. 4b) as an EDA shows that the data actually support quite a number of possible roots. These are, in approximate order of decreasing split support (as indicated by the edge lengths), the edges leading to: (1) the grasses, (2) *Amborella* + *Nymphaea*, (3) *Amborella*, (4) *Nymphaea*, (5) *Calycanthus*, (6) the grasses + other monocots minus *Acorus*, (7) *Amborella* + *Calycanthus*, (8) *Amborella* + *Calycanthus* + *Nymphaea*, and (9) *Acorus* + *Amborella* + *Nymphaea*. So, the best-supported root is the one shown in the neighbor-joining tree, as expected, but rooting somewhere along edge c is also well supported. Edge d, however, is firmly rejected by the data as a possible root (and the monocotyledons must therefore be paraphyletic).

Holland et al. (2006) provide a somewhat different EDA of this dataset. They first constructed a separate maximum-likelihood tree for each of the 61 genes, and then calculated the consensus network of these trees, filtering out the least-supported bipartitions in the process. Their data-display network shows support (in approximate order) for roots (2), (1), (3) and (4). This is a somewhat slower route (i.e. more calculation-intensive) to roughly the same conclusion as I have reached here.

Some other examples of methodological problems identified using data-display networks are provided by Morrison (2010), including sequence variation between laboratories and sequence misalignment, along with examples of homoplasy due to parallelism and reversal.

Detecting Reticulation Processes

In addition to homoplasy and methodological issues, it is also possible for EDA to be used to identify possible gene-flow processes, which may or may not be the main focus of the phylogenetic

study. However, it is best to remember that even under these circumstances the data-display networks should not be interpreted as evolutionary networks, because the nodes do not necessarily represent ancestors and the edges do not necessarily represent biological events. Data-display networks are good for *generating* biological hypotheses but not for *testing* them. Here, I illustrate the use of data-display networks to investigate recombination, hybridization, introgression and incomplete lineage-sorting.

The splits network in Fig. 5 has a single "netted region" (a collection of cycles with shared edges), apparently involving one of the *Rana palmipes* samples (labeled VenAMNHA118801). This sample shares slightly more similarity with the other *R. palmipes* sample (EcuKU204425) than it does with the *R. spectabilis* sample (JAC8622). Cross-checking with the original sequence data shows that the first 3/5 of the alignment is shared between the two *R. palmipes* samples while the final 2/5 is shared with *R. spectabilis*. If the idea of a sequencing error (e.g. *in vitro* recombination) is rejected, then perhaps the most likely explanation here is recombination, even though these are mitochondrial sequences collected quite some geographical distance apart. The original authors noted that the VenAMNHA118801 sample was "considerably more divergent" but did not note the specific pattern in the sequence. They did, however, suggest that there might be multiple species in *R. palmipes*. Irrespective of its cause, this "recombination" pattern might confound any subsequent tree-building analysis. In the authors' maximum-likelihood tree, the two *R. palmipes* samples are sisters but the VenAMNHA118801 sample has a very long terminal edge, while in the parsimony analysis the latter sample is placed elsewhere in the tree.

The splits network in Fig. 6 has two cycles representing character conflict. The reticulation involving *Viburnum prunifolium* is caused by conflict between the two genes, rather than by conflict within the genes. This was identified by the original authors as resulting from a hybridization of *V. lentago* and *V. rufidulum*, which is certainly consistent with the data-display network. However, the involvement of *V. elatum* is also a possibility based on the diagram, due to

the polychotomy, or even the ancestor of *V. rufidulum* and *V. elatum*. The other reticulation, involving *V. melanocarpum*, is caused by conflict within the genes, which was not discussed by the original authors. It is unlikely to be a hybridization event, because it would involve *V. erosum* and the ancestor of *V. erosum* and *V. japonicum*, which would not be time-consistent (i.e. a descendant would be hybridizing with its own ancestor). It is probably a straightforward case of homoplasy. It does, however, cause lack of resolution in the authors' tree-building analyses.

The next example involves an EDA of a potential case of introgression due to complex sharing of mitochondrial and nuclear DNA patterns. The original author compared parsimony trees from both nuclear and mitochondrial sequences, concluding from their conflict that there was "introgression of *Drosophila simulans* III mtDNA into *D. mauritania*." For the nuclear-DNA analysis, the two *D. mauritania* samples were sisters in the parsimony consensus tree, as were samples I and III from *D. simulans*, with 97% and 62% bootstrap support, respectively. These two placements are crucial to the author's argument. However, Fig. 7 shows that neither of these placements is uncontradicted by the character data, as indicated by the conflicting bipartitions in the parsimony splits network. More importantly, the support for the *D. mauritania* placement comes almost entirely from the choice of a parsimony tree-building analysis. A split decomposition analysis using the simplest model that corrects for multiple substitutions (the Jukes-Cantor model) shows that most of the support for this sister-group relationship disappears, although the *D. simulans* sister-group relationship remains (albeit with much character conflict). This greatly weakens the strength of the author's conclusions.

It is important to note that in this example the network patterns do not reflect the tree bootstrap supports, and the strongest bootstrap support in the tree is associated with the weakest pattern in the network. This independence of EDA information and bootstrapping has also been noted by Wägele and Mayer (2007). EDA provides information prior to tree building whereas bootstrapping provides it afterwards.

Moving on, the splits network in Fig. 8 has two netted regions and two simple cycles representing character conflict. All of the patterns shown here as reticulations were attributed by the original authors to incomplete lineage-sorting (or ancestral polymorphism). This is because of the sharing of haplotypes between the different *Senecio* species (notably *S. glaucus*) and the complex linking of haplotypes between species (notably in *S. leucanthemifolius*, *S. glaucus* and *S. rupestris*). This interpretation is certainly consistent with the data-display network. However, the authors reach their conclusions based on a series of separate maximum-parsimony trees (one for each data type), which is a much more complicated way of displaying the same data conflicts shown here.

Similarly, in Fig. 9 the pattern shown by the single netted region was attributed by the original authors to incomplete lineage-sorting. However, they presented a very simplified phylogenetic tree (of unclear origin), which does not make obvious the complex patterns of sharing of haplotypes between the genera. A network seems to be a much superior display of the relevant data here.

For the example in Fig. 10, the original authors were interested in the two rDNA forms that they detected in *Perkinsus andrewsi* (labeled A and B). However, the EDA makes it clear that these forms differ in four apparently independent ways, represented by the four numbered splits in Fig. 10a. These sequence differences are shared with sequences from the other three species, rather than being within-species divergences. Furthermore, the character support for these splits is not randomly distributed along the sequenced region, but occurs at distinct non-overlapping places within the ITS1 and ITS2 genes (Fig. 10b). Of particular interest, in the sequences labeled “sp.2c”, “sp.1c” and “sp.2a” the rDNA unit is apparently a mosaic, each one containing some but not all of the differences apparent between the two *P. andrewsi* forms. Clearly, this situation has not arisen by simple vertical inheritance. The original authors suggest *in vitro* recombination as a possible cause, but it could be a gene duplication followed by parallel changes among the species, or some form of introgression. Whatever the explanation, the circumstances are much more complex than could be

resolved by a phylogenetic tree alone.

Unresolved Patterns

In this section I illustrate the use of data-display networks to provide an EDA of whether there is, indeed, sufficient unconflicted information in a dataset for it to be worthwhile to perform a phylogenetic analysis. A tree, after all, summarizes the majority data pattern in the characters, and if there is no clear majority then a tree will not be a valid representation of that dataset. An EDA seems to be an obvious step to take before starting any tree building, and yet there is almost no indication in published papers that it is performed.

The splits network in Fig. 11 has a single cycle. The long internal edges indicate strong support for the four main groups of *Leishmania* (subgenera *Sauroleishmania* and *Viannia*, and two clades within subgenus *Leishmania*). However, the short central edges forming the cycle indicate two conflicting signals, so that the relationships between the subgenera are not clear. There are simply too few parsimony-informative characters to resolve the short edges and make an unequivocal decision about the relationships of three of the groups (there are 6 versus 5 characters supporting the two resolutions shown in the reticulation). This conflict is not represented in the parsimony tree, as there is 80% bootstrap support for the clade *Leishmania* + *Sauroleishmania*. A tree is probably valid here, but it should bear one notable unresolved polychotomy.

A more severe situation is where there are few data about relationships among many of the samples. The example in Fig. 12 shows a splits network with apparent support for only three pairs out of the 10 samples, and even then this support is not uncontradicted. Otherwise, there are simply 10 long terminal edges, with very little in the way of information about the inter-relationships. The EDA shows that much of the sequence data (10 768 aligned nucleotides) are either invariant (5944 positions) or relate to autapomorphies (1524 positions), with only 31% of the data potentially

relating to shared relationships. Furthermore, these synapomorphic data are very contradictory, with 47 of the possible 255 bipartitions (for 10 taxa) having at least some data support. Tree building will not be a reliable activity under these circumstances.

An even more extreme example is discussed by Wägele and Mayer (2007) (their Fig. 20). Here, the synapomorphic data were almost completely contradictory, with *all* of the possible 127 bipartitions (for 8 taxa) having at least some data support. Not unexpectedly, the EDA network was simply a spider web with 8 long edges sticking out of it. Notwithstanding this, the bootstrap analysis of the original authors showed 91–100% support for 3 of the internal edges of their phylogenetic tree. It is difficult to see what value a tree has in this situation.

In population datasets it is more common to encounter network analyses compared to datasets dealing with species, but many such data sets are still analyzed by tree building alone. Fig. 13 shows both a tree analysis and a data-display network analysis of the same dataset. The neighbor-net network is based on the neighbor-joining tree, and this is clearly reflected in the underlying "skeleton" of the network graph. However, the true population complexity of character relationships is obscured in the tree alone, which therefore gives a rather false visual impression of the population patterns.

Failure of the Analysis

Here, I illustrate the use of data-display networks to investigate the failure of an evolutionary network to analyze the data correctly, thus illustrating one of the inherent limitations of the current algorithms.

One method for producing an evolutionary network is to "simplify" the cycles of a data-display network, thereby possibly filtering the stochastic data conflicts from the biologically relevant ones. Huson and Klöpper (2007) point out that a splits network has a set of independent

reticulations that form "netted regions" while a "reticulate network" has a set of dependent reticulations that form "tangles" (or galls). Tangles are simpler than netted regions but nevertheless preserve relationships among the samples. So, a splits network (or any collection of bipartitions) can be converted into a reticulate network by taking the unrooted graph and adding a root, and then finding the minimum number of reticulation nodes needed to replace each netted region with a tangle. The objective is to find compatible reticulations that can be pooled, reducing all of the data incompatibilities to simpler reticulation events (McBreen and Lockhart 2006).

The example in Fig. 14 illustrates both the method and its limitation. The top graph shows a "known" species tree with four speciation events, plus two horizontal gene transfers (HGTs) affecting taxa C+A and B+F. In this example, each HGT involves one gene only, so there are two resulting gene trees. These trees have no components (or non-trivial bipartitions) in common. Nevertheless, if each gene tree is reconciled with the rooted species tree, using the program of Addario-Berry et al. (2003), then the reticulate taxon is correctly identified in each case (i.e. taxa A and F). Note that I am using the term "reticulate taxon" because there is assumed to be no *a posteriori* knowledge about whether the gene-flow process is hybridization, recombination or HGT.

A consensus network (Holland et al. 2006) can also be constructed from these two gene trees, which is a type of data-display network. This would be suitable for an EDA; and it indicates an apparently complex set of inter-relationships among the taxa. Sadly, we know that this EDA is misleading, since there are only two instances of gene flow that are independent of the normal divergent process of vertical inheritance. This unfortunate situation occurs because neither of the gene trees reflects the underlying species tree (i.e. the tree for the rest of the genome).

We could then also construct a reticulate network (Fig. 14 bottom) from the consensus network, with taxon A providing the root. In this case, it would be a "hybridization network" since the underlying mathematical model is based on the biological process of hybridization. This is now an evolutionary network (rather than a data-display network), with each edge having a direction

away from the root. It can be interpreted in terms of ancestors and descendants, with each edge representing a biological event transforming an ancestor into one of its descendants. This network identifies taxa D and E as the reticulate ones. Note that these are the only two taxa that are *not* involved in the HGTs!

The issue with this artificial example is that there are no data relating directly to the tree-like pattern of vertical inheritance. That is, I have not included a gene tree that is unaffected by HGT. Clearly, reconstructing evolutionary networks in the absence of data relating directly to vertical inheritance is impractical. This appears to be a fundamental limitation of any method for constructing evolutionary networks, rather than something that is specific to the particular algorithm used in my example.

Conclusion

In each of the 13 empirical datasets that I have examined here (plus one hypothetical example), there is at least one important point (relating to conflicting data patterns) that will affect the phylogenetic analysis. These effects will not always result in mis-interpretation of the results, but they will certainly influence the perceived strength of the support for the final inferences.

I therefore conclude that EDA should play a greater role in phylogenetic analyses than it has done to date. It should be seen as an essential first step in the analysis of any dataset that is to be used for phylogenetic analysis. In particular, it is essential to remember that a tree-building analysis fits the data to a tree irrespective of whether the data are tree-like or not. This fundamental assumption should be assessed prior to any attempt to force the data into a tree.

Also, it is important to understand the difference between data-display networks, which are useful for EDA, and evolutionary networks, which generalize phylogenetic trees. The challenge of reconstructing networks that explicitly represent the evolution of taxa, rather than merely displaying

data conflict, is being met by several research groups. Whether this is a practical goal for most realistic data sets is an open question.

Acknowledgments

Thanks to the various authors for making their datasets available, and to the referees for some very helpful suggestions.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 21:2104–2105.

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

Ayling SC, Brown TA. 2008. Novel methodology for construction and pruning of quasi-median networks. BMC Bioinform. 9:115.

Ballard JWO. 2000. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. Mol Biol Evol. 17:1126–1130.

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

Bandelt H-J. 2005. Exploring reticulate patterns in DNA sequence data. In: Bakker FT, Chatrou LW, Gravendeel B, Pelser PB, editors. Plant species-level systematics: new perspectives on pattern and process. Königstein: Koeltz. pp. 245–269.

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

Bandelt H-J, Dress AWM. 1993. A relational approach to split decomposition. In: Opitz O, Lausen

B, Klar R, editors. Information and classification. Berlin: Springer. pp. 123–131.

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

Behrens JT, Yu CH. 2003. Exploratory data analysis. In: Schinka JA, Velicer WF, editors. Handbook of psychology, vol. 2: research methods in psychology. Hoboken NJ: John Wiley & Sons. pp. 33–64.

Bergsten J. 2005. A review of long-branch attraction. Cladistics. 21:163–193.

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

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

Comes HP, Abbott RJ. 2001. Molecular phylogeography, reticulation, and lineage sorting in Mediterranean *Senecio* sect. *Senecio* (Asteraceae). Evolution. 55:1943–1962.

Cooper AS., Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R. 2001. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature. 409:704–707.

Croan DG, Morrison DA, Ellis JT. 1997. Evolution of the genus *Leishmania* revealed by comparison of DNA and RNA polymerase gene sequences. Mol Biochem Parasitol. 89:149–159.

Donoghue MJ, Baldwin BG, Li J, Winkworth RC. 2004. *Viburnum* phylogeny based on chloroplast *trn*K intron and nuclear ribosomal ITS DNA sequences. Syst Bot. 29:188–198.

Ellison AM. 2001. Exploratory data analysis and graphic display. In: Scheiner SM, Gurevitch J, editors. Design and analysis of ecological experiments, 2nd edn. Oxford: Oxford Uni. Press. pp. 37–62.

Gemeinholzer B. 2008. Phylogenetic networks. In: Junker BH, Schreiber F, editors. Analysis of biological networks. Hoboken NJ: Wiley-Interscience. pp. 255–282.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol. 22:1813–1822.

Grant T, Kluge AG. 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics. 19:379–418.

Hillis DM, Wilcox TP. 2005. Phylogeny of the New World true frogs (*Rana*). Mol Phylogenet Evol. 34:299–314.

Höglund J, Morrison DA, Mattsson JG, Engström A. 2006. Population genetics of the bovine/cattle lungworm (*Dictyocaulus viviparus*) based on mtDNA and AFLP marker techniques. Parasitology. 133:89–99.

Holland BR, Delsuc F, Moulton V. 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. Syst Biol. 54:66–76.

Holland BR, Huber KT, Moulton V, Lockhart PJ. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol. 21:1459–1461.

Holland BR, Jermiin LS, Moulton V. 2006. Improved consensus network techniques for genome-scale phylogeny. Mol Biol Evol. 23:848–855.

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

Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 14:68–73.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Huson DH, Dezulian T, Klöpper T, Steel MA. 2004. Phylogenetic super-networks from partial

trees. IEEE/ACM Trans Computat Biol Bioinform. 1:151–158.

Huson DH, Kloepper TH. 2005. Computing recombination networks from binary sequences. Bioinformatics. 21:ii159–ii165.

Huson DH, Klöpper TH. 2007. Beyond galled trees – decomposition and computation of galled networks. Lect Notes Bioinform. 4453:211–225.

Huson DH, Klöpper T, Lockhart PJ, Steel MA. 2005. Reconstruction of reticulate networks from gene trees. Lect Notes Bioinform. 3500:233–249.

Huson DH, Rupp R, Berry V, Gambette P, Paul C. 2009. Computing galled networks from real data. Bioinformatics. 25:i85–i93.

Huson DH, Steel MA, Whitfield J. 2006. Reducing distortion in phylogenetic networks. Lect Notes Bioinform. 4175:150–161.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol 53:638–643.

Kennedy M, Holland BR, Gray RD, Spencer HG. 2005. Untangling long branches: identifying conflicting phylogenetic signals using spectral analysis, neighbor-net, and consensus networks. Syst Biol. 54:620–633.

Makarenkov V, Kevorkov D, Legendre P. 2006. Phylogenetic network construction approaches. In: Arora DK, Berka RM, Singh GB, editors. Applied mycology and biotechnology, vol. 6: bioinformatics. Amsterdam: Elsevier. pp. 61–97.

McBreen K, Lockhart PJ. 2006. Reconstructing reticulate evolutionary histories of plants. Trends Plant Sci. 11:398–404.

Morrison DA. 1996. Phylogenetic tree-building. Int J Parasitol. 26:589–617.

Morrison DA. 2005. Networks in phylogenetic analysis: new tools for population biology. Int J Parasitol. 35:567–582.

Morrison, DA. 2010. Phylogenetic networks in systematic biology (and elsewhere). In Mohan RM, editor. Research advances in systematic biology. Trivandrum, India: Global Research Network.

Nakhleh L. 2010. Evolutionary phylogenetic networks: models and issues. In: Heath LS, Ramakrishnan N, editors. The problem solving handbook for computational biology and bioinformatics. New York: Springer.

Nakhleh L, Sun J, Warnow T, Linder CR, Moret BME., Tholse A. 2003. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. Pacific Symp Biocomp. 8:315–326.

Pecher WT, Robledo JAF, Vasta GR. 2004. Identification of a second rRNA gene unit in the *Perkinsus andrewsi* genome. J Eukaryot Microbiol. 51:234–245.

Philip GK, Creevey CJ, McInerney JO. 2005. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi, and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol. 22:1175–1184.

Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol. 6:498–505.

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

Reeves PA, Richards CM. 2007. Distinguishing terminal monophyletic groups from reticulate taxa: performance of phenetic, tree-based, and network procedures. Syst Biol. 56:302–320.

Takahashi K, Terai Y, Nishida M, Okada N. 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Mol Biol Evol. 18:2057–2066.

Troell K, Engström A, Morrison DA, Mattsson JG, Höglund J. 2006. Global patterns reveal strong population structure in *Haemonchus contortus*, a nematode parasite of domesticated

ruminants. Int J Parasitol. 36:1305–1316.

Tukey JW. 1977. Exploratory data analysis. Reading MA: Addison-Wesley.

Vriesendorp B, Bakker FT. 2005. Reconstructing patterns of reticulate evolution in angiosperms: what can we do? Taxon. 54:593–604.

Wägele JW, Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol Biol. 7:147.

Table 1

Phenotypic Data Matrix for Some Extant Vertebrates

Taxon	Characters 1–16 [a]
Lampreys	1000000000000000
Sharks	1100000000000001
Teleosts	1110000000000002
Lungfishes	1111100000000002
Frogs	1111111000000003
Salamanders	1111111000000003
Turtles	1111110100000004
Lizards	1111110111100004
Snakes	1111110111100004
Crocodiles	1111110110011104
Birds	1111110110011115
Mammals	1111110120000016

[a] The data are from Morrison (1996).

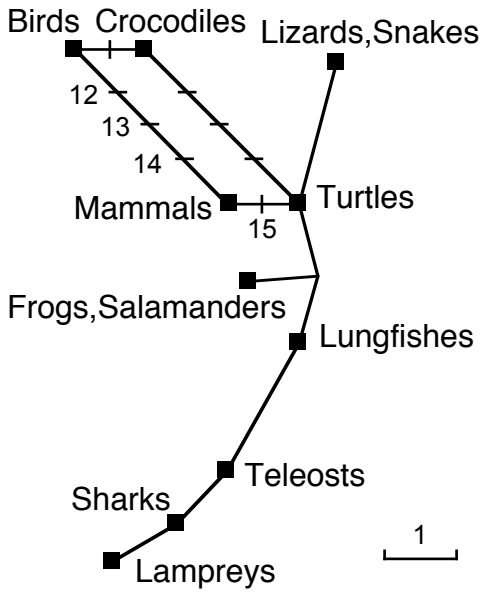

FIG. 1.—Parsimony splits analysis of the morphological data matrix in Table 1. The scale bar represents the number of character-state changes (non-binary characters have been ignored by the analysis). There are no autapomorphies, so most of the taxa appear as internal nodes; and two pairs of the taxa are identical for the analyzed data and so are plotted at the same location.

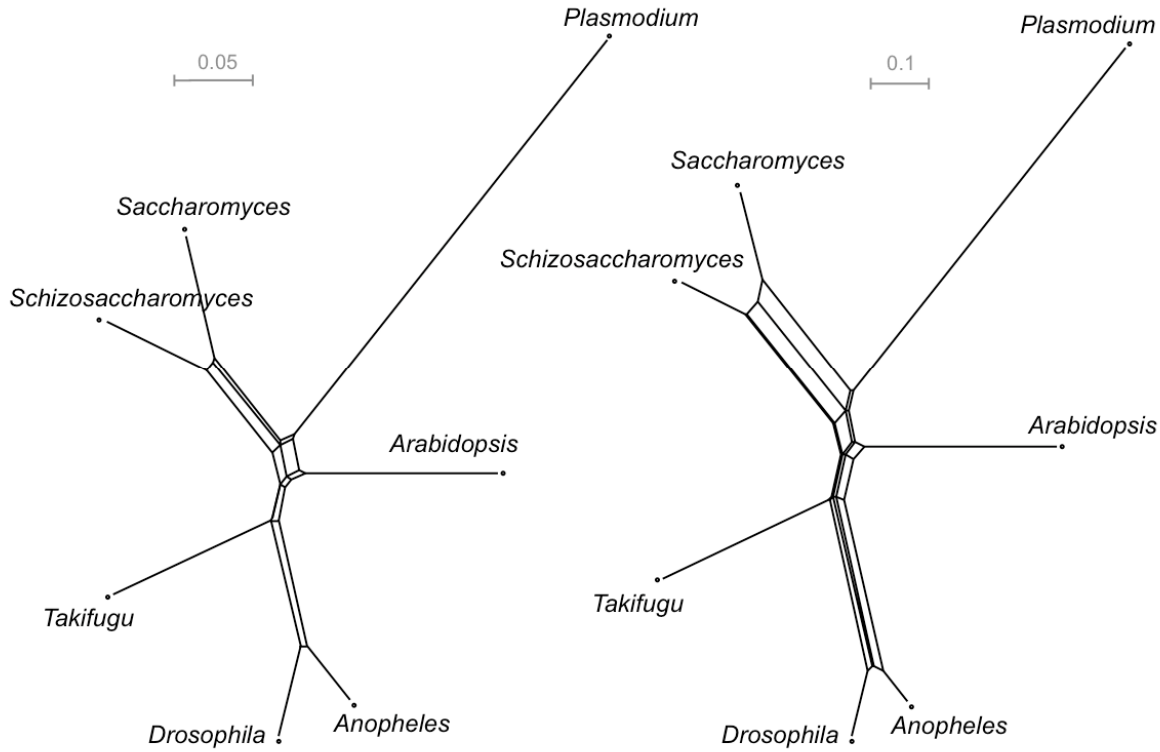

FIG. 2.—Comparison of two neighbor-net networks, based on different amino-acid substitution models for 3579 aligned amino acids from 7 eukaryote taxa. The scale bar represents the split support for the edges. The data are sequences of an unidentified gene named 6471 from Philip et al. (2005). The left graph shows use of the WAG+G substitution matrix, while the right graph shows use of the MtMAM+G matrix. The two analyses were otherwise identical, so that the difference in network complexity is due entirely to the differing substitution models.

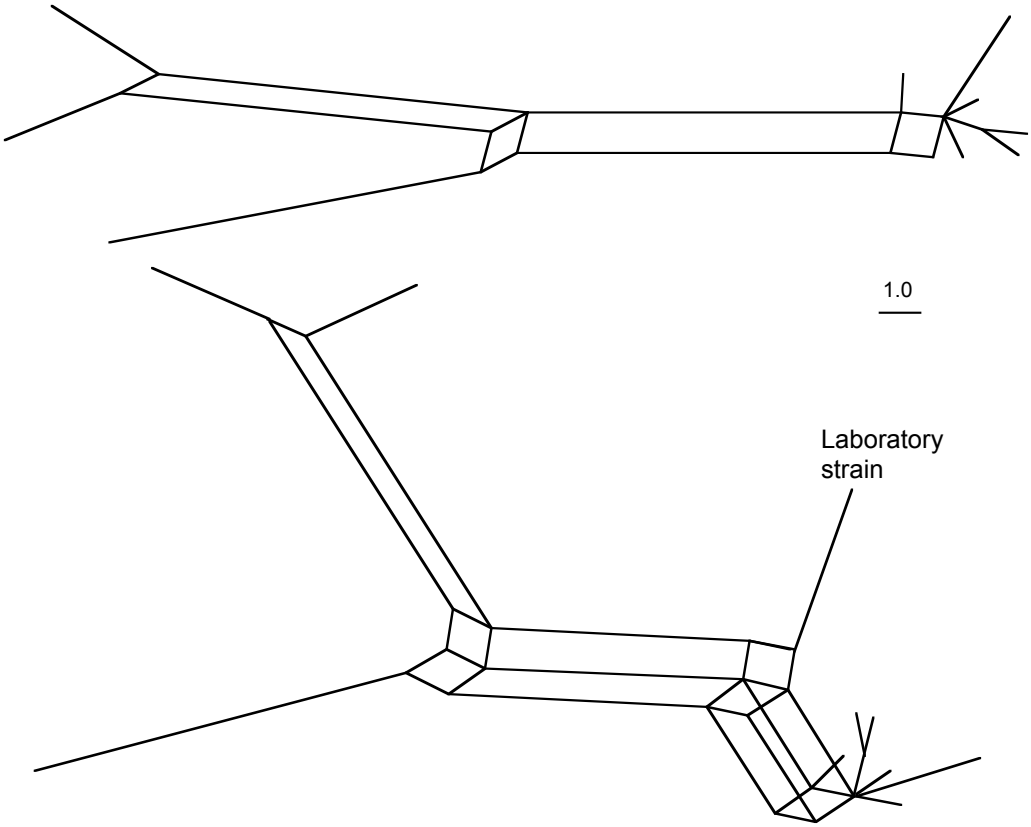

FIG. 3.—Comparison of two median networks, with and without an outgroup, for 1542 aligned nucleotides from 64 farm samples of *Dictyocaulus viviparus* (Nematoda). The networks are based on binary characters only; and there are many identical haplotypes (which are plotted at the same location). The scale bar represents one character-state change. The data are mitochondrial protein, rRNA and tRNA gene sequences from Höglund et al. (2006). The lower analysis differed from the top one solely in the addition of 8 samples from a laboratory strain (as a potential outgroup).

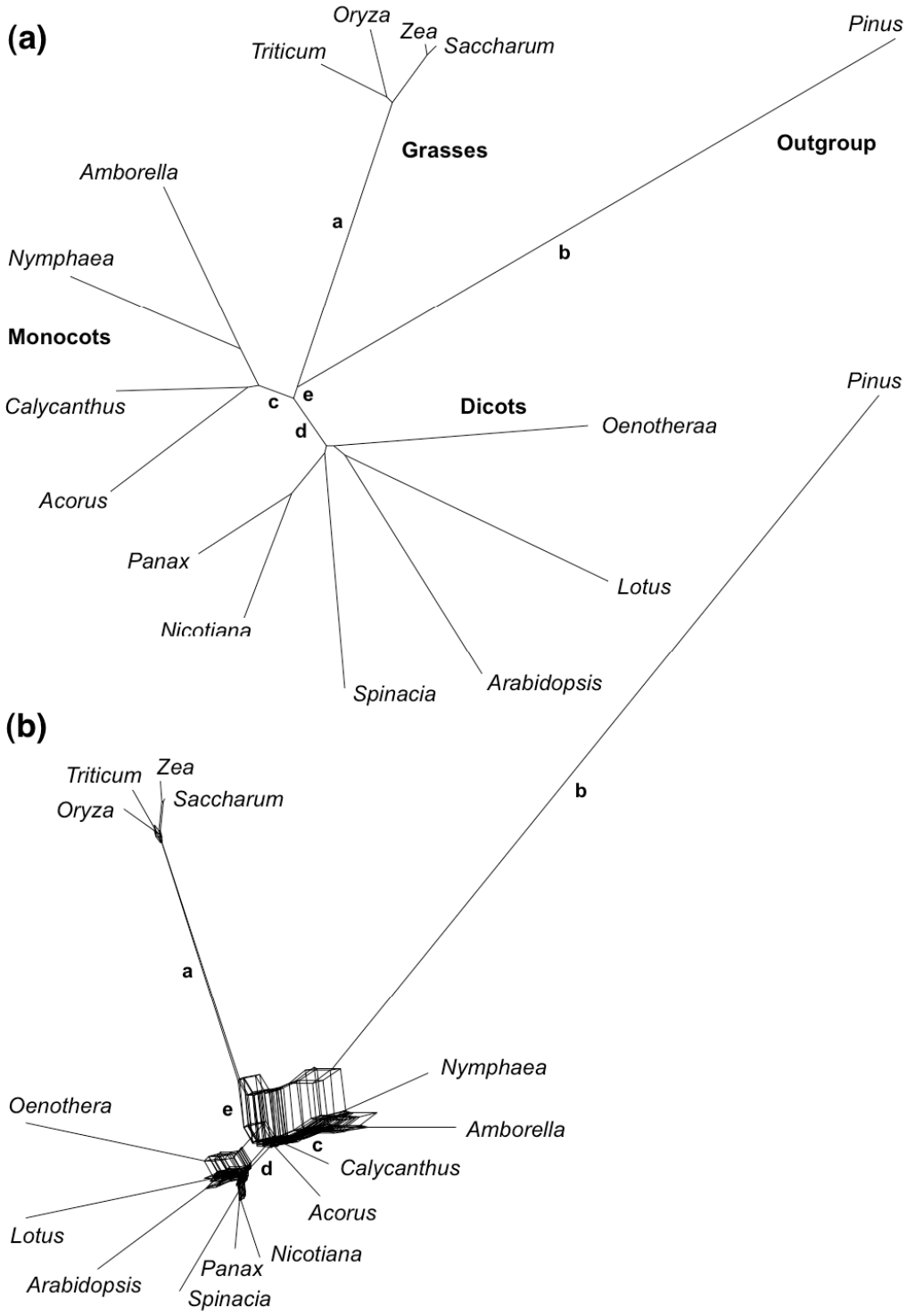

FIG. 4.—Comparison of (a) the neighbor-joining tree (based on the hamming distance) and (b) the median network (filtered to a minimum support of 4) for 89 436 aligned nucleotides from 15 plant species. The edge lengths represent (a) inferred evolutionary change and (b) split support for the

edges. The five edges possibly involved in long-branch attraction are lettered. The data are 61 chloroplast gene sequences from Goremykin et al. (2005).

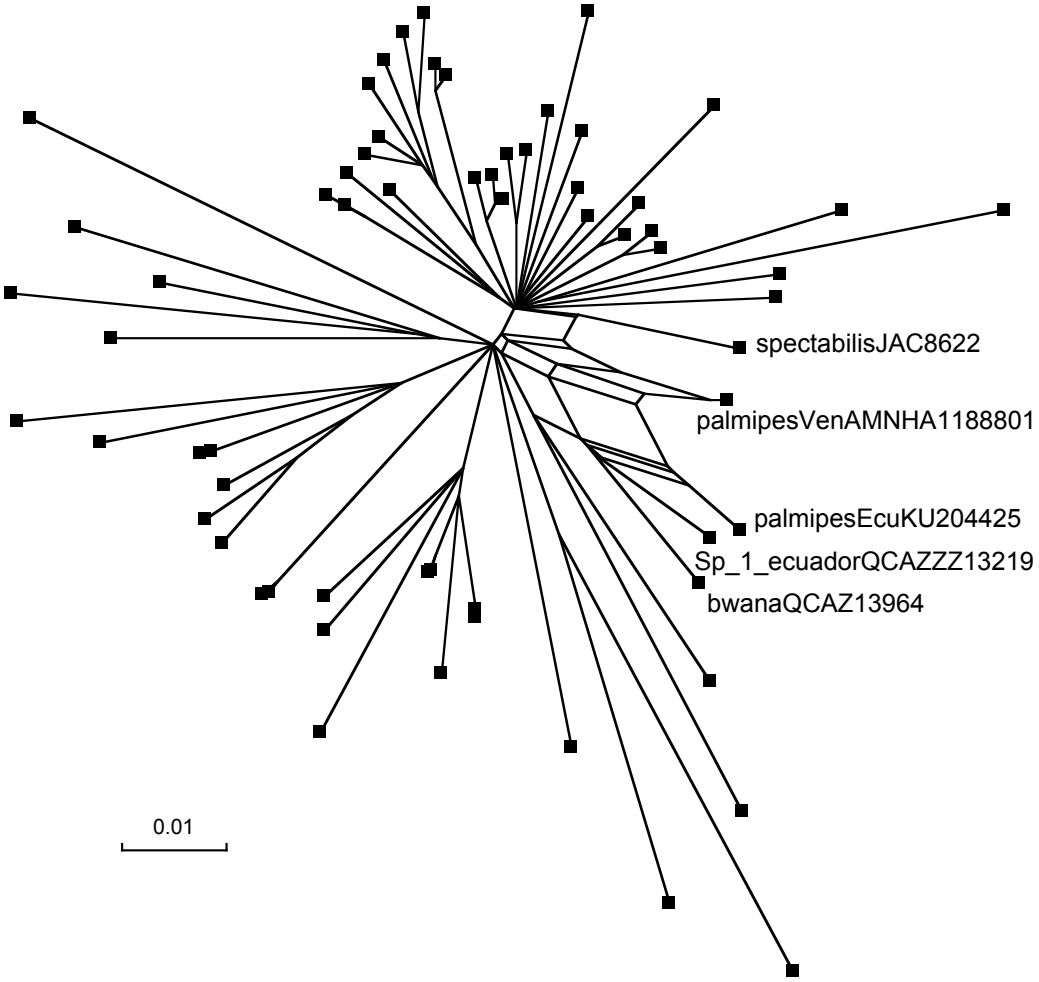

FIG. 5.—Split decomposition analysis (based on the hamming distance) for 1976 aligned nucleotides from 64 *Rana* samples (Amphibia). Most of the samples are unlabeled; and the scale bar represents the split support for the edges. The data are mitochondrial rRNA and tRNA sequences from Hillis and Wilcox (2005). The relative lengths of the edges in the netted region show that there is slightly more support for the grouping of palmipesVenAMNHA118801 with palmipesEcuKU204425 than with spectabilisJAC8622.

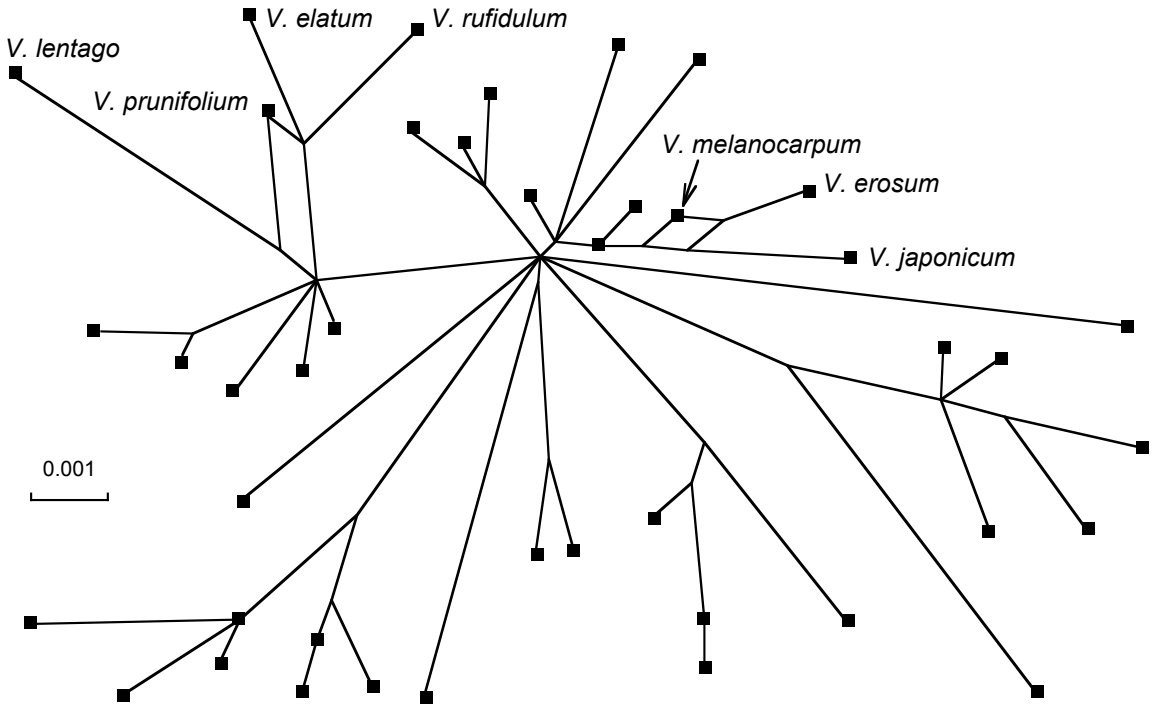

FIG. 6.—Split decomposition analysis (based on the hamming distance) for 1687 aligned nucleotides from 45 *Viburnum* samples (Plantae). Most of the samples are unlabeled; and the scale bar represents the split support for the edges. The data are chloroplast ITS and *trn*K sequences from Donoghue et al. (2004). There are two undirected cycles.

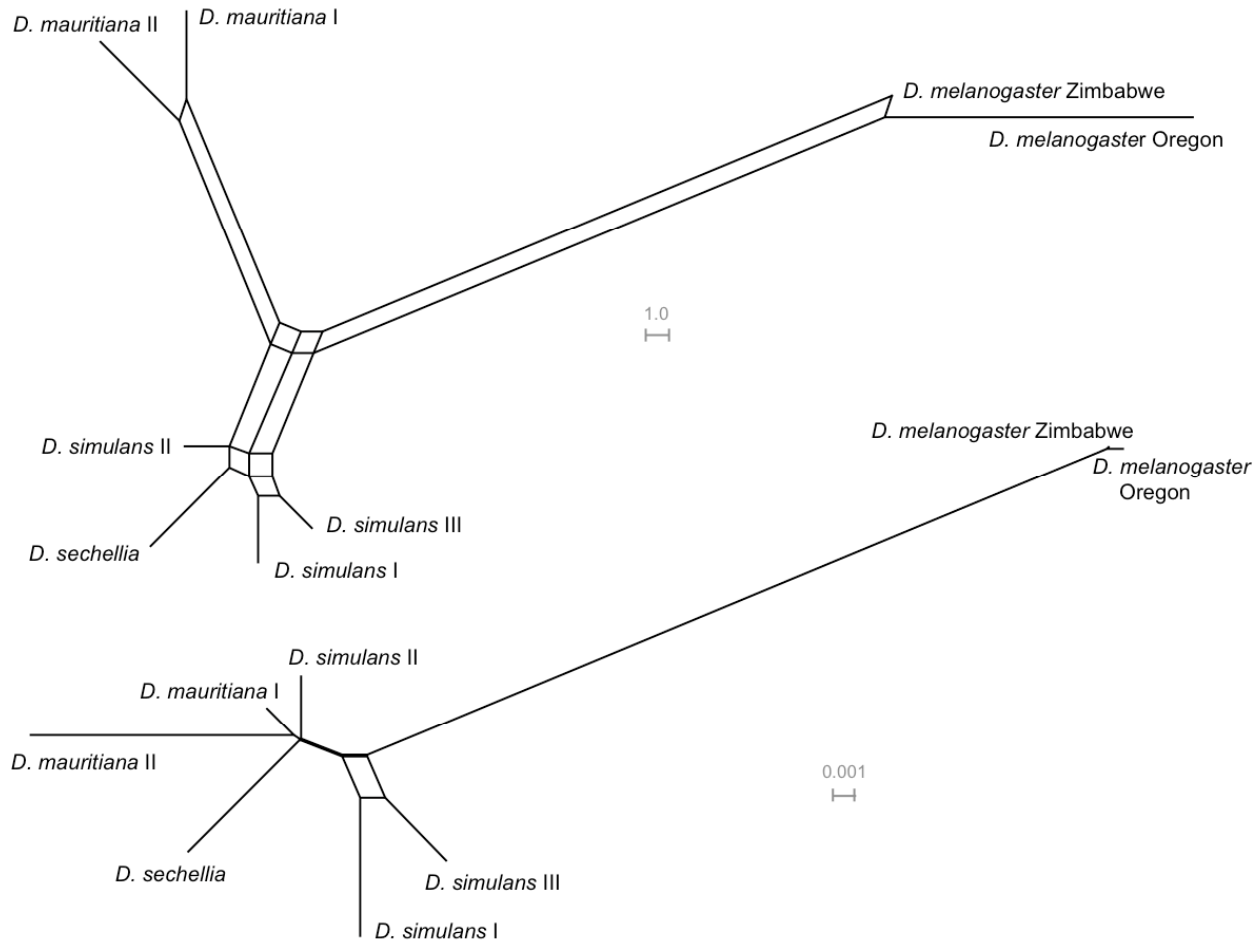

FIG. 7.—Comparison of parsimony splits (above) and split decomposition (below) analyses for 501 aligned nucleotides from 8 *Drosophila* samples (Insecta). The scale bar represents the number of character-state changes (above) or the split support for the edges (below). The data are nuclear *Adhr* sequences from Ballard (2000). The splits graphs differ in that the split decomposition used the JC-corrected distance while the parsimony splits used the original characters.

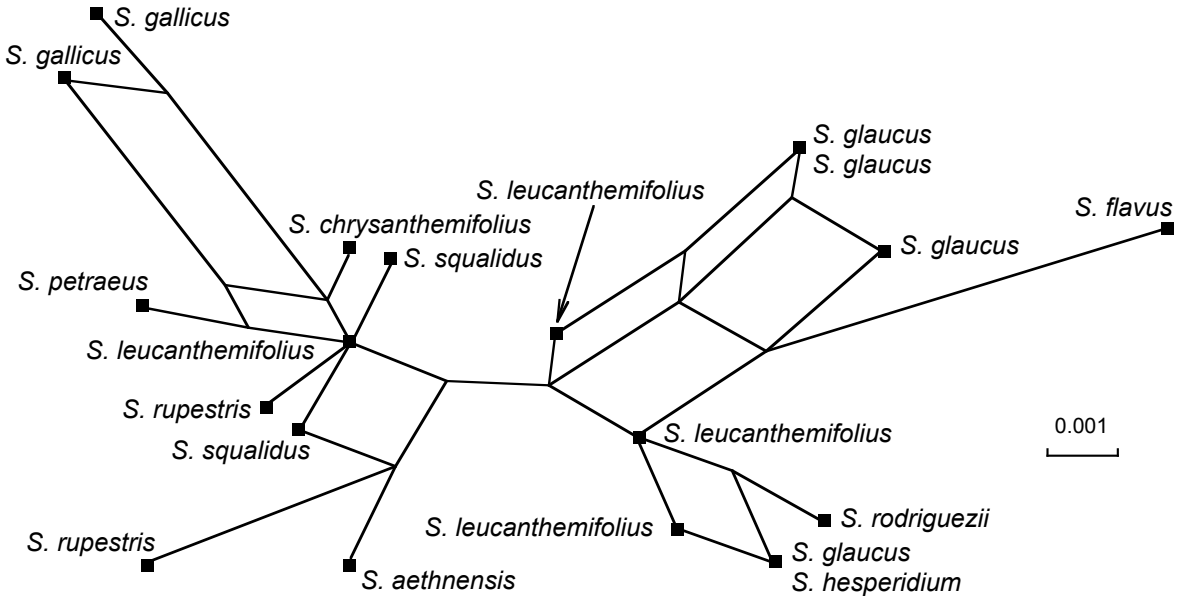

FIG. 8.—Split decomposition analysis (based on the hamming distance) for haplotypes from 21 *Senecio* samples (Plantae). Note that several of the species names appear multiple times (representing different samples); and the scale bar represents the split support for the edges. The data are 649 aligned nuclear ITS positions plus cpDNA RFLP haplotypes from Comes and Abbott (2001). There are two netted regions and two simple cycles.

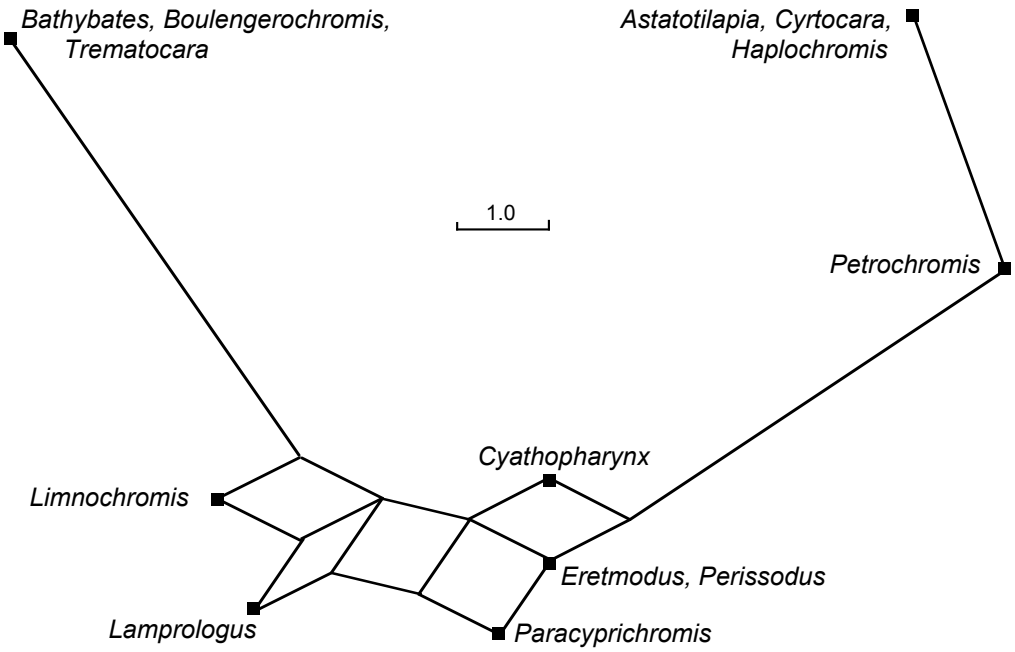

FIG. 9.—Median network for 20 SINEs from 13 cichlid taxa (Pisces). There are several identical haplotypes (which are plotted at the same location); and the scale bar represents one character-state change. The data are short interspersed nuclear element (SINE) loci from Takahashi et al. (2001). There is a single netted region.

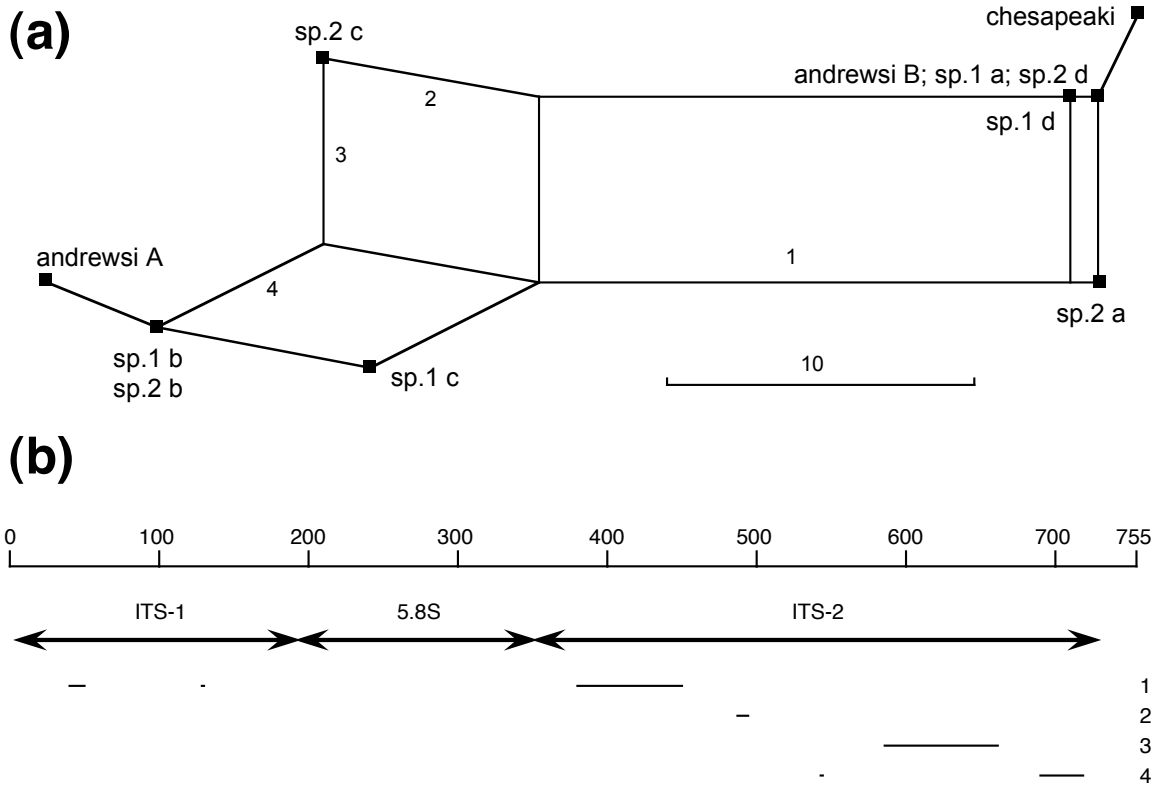

FIG. 10.—(a) Parsimony splits analysis of 755 aligned nucleotides from 11 samples for 4 *Perkinsus* species (Alveolata). There are several identical haplotypes (which are plotted at the same location); and the scale bar represents the number of character-state changes. Four of the 5 splits are numbered (and discussed in the text). The data are nuclear rRNA sequences from Pecher et al. (2004). (b) Schematic representation of the sequenced part of the rRNA locus, showing the parts of the sequence that support the four splits numbered in part (a). Split 1 = 18 characters (positions 40–48, 378–450) + 1 not shared by sp.1_d (at position 130); Split 2 = 7 characters in one indel (positions 488–494); Split 3 = 7 characters (positions 586–662); and Split 4 = 6 characters (positions 545, 691–718).

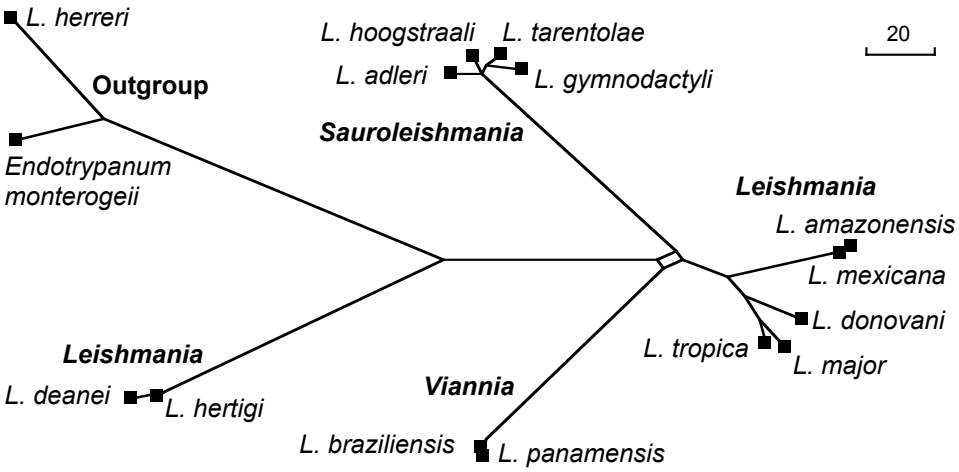

FIG. 11.—Parsimony splits analysis of 2207 aligned nucleotides from 15 *Leishmania* species (Kinetoplastida). The subgenera are labelled; and the scale bar represents the number of character-state changes. The data are nuclear DNA and RNA polymerase sequences from Croan et al. (1997). There is a single undirected cycle.

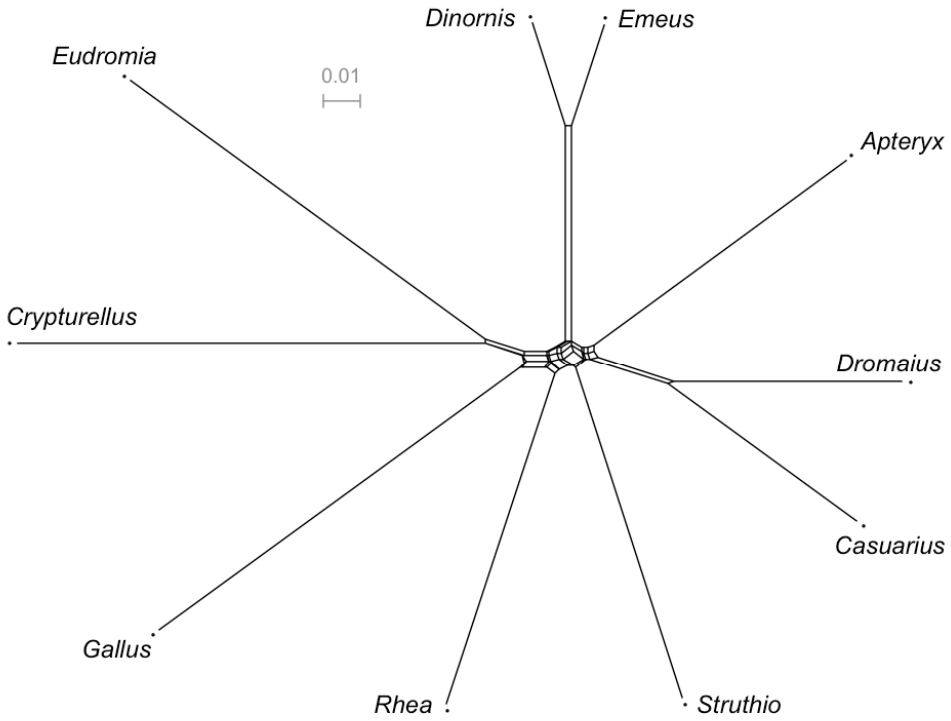

FIG. 12.—Neighbor-net analysis (based on the hamming distance) of 10 768 aligned nucleotides from 10 species of ratites (Aves). The scale bar represents the split support for the edges. The data are 12 mitochondrial protein-gene sequences from Cooper et al. (2001). There is a large netted region of undirected cycles.

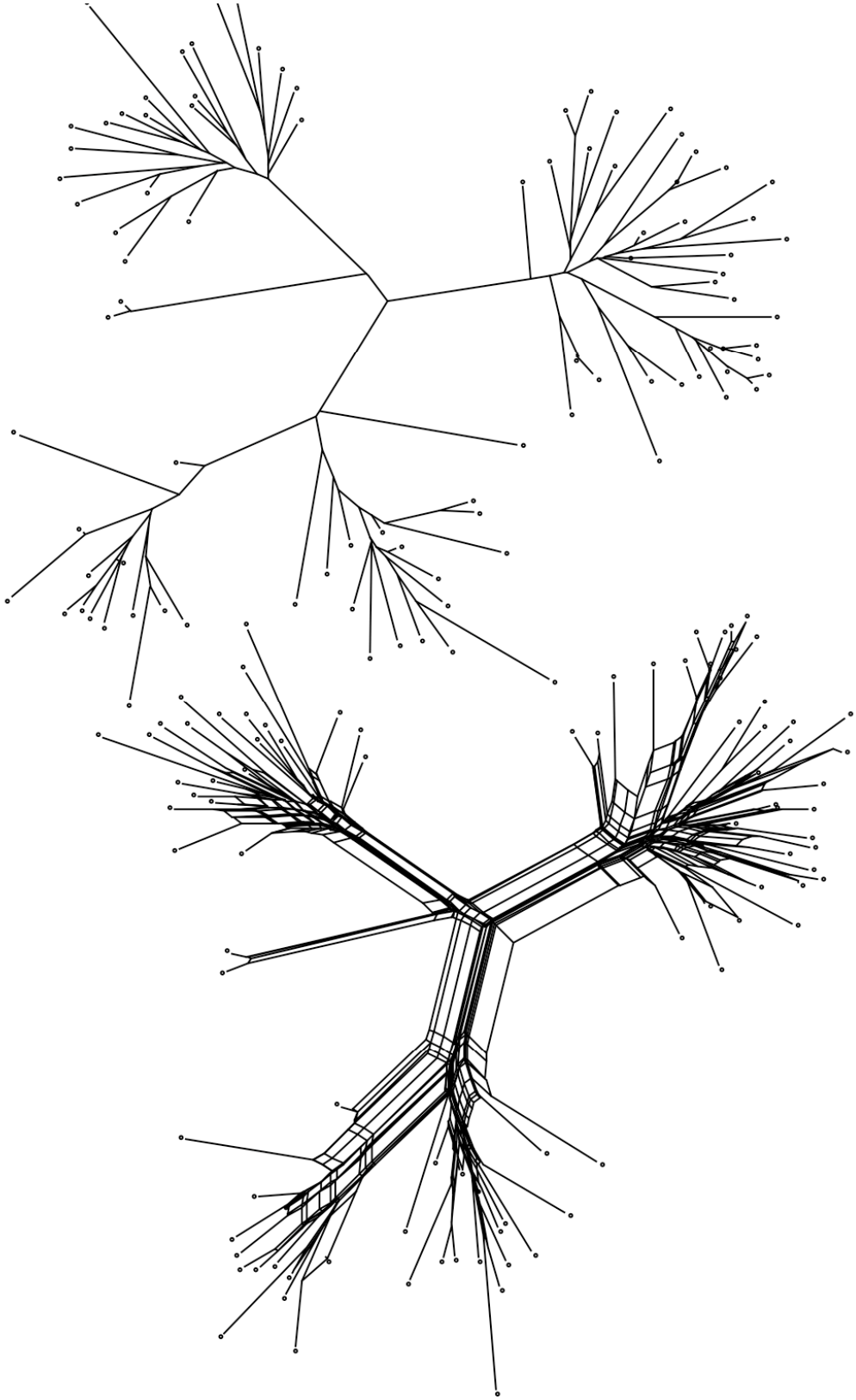

FIG. 13.—Comparison of neighbor-joining tree (above) and neighbor-net network (below) for 411 aligned nucleotides from 94 samples of *Haemonchus contortus* (Nematoda). Both analyses are

based on the hamming distance. The data are mitochondrial *nad*4 sequences from Troell et al. (2006).

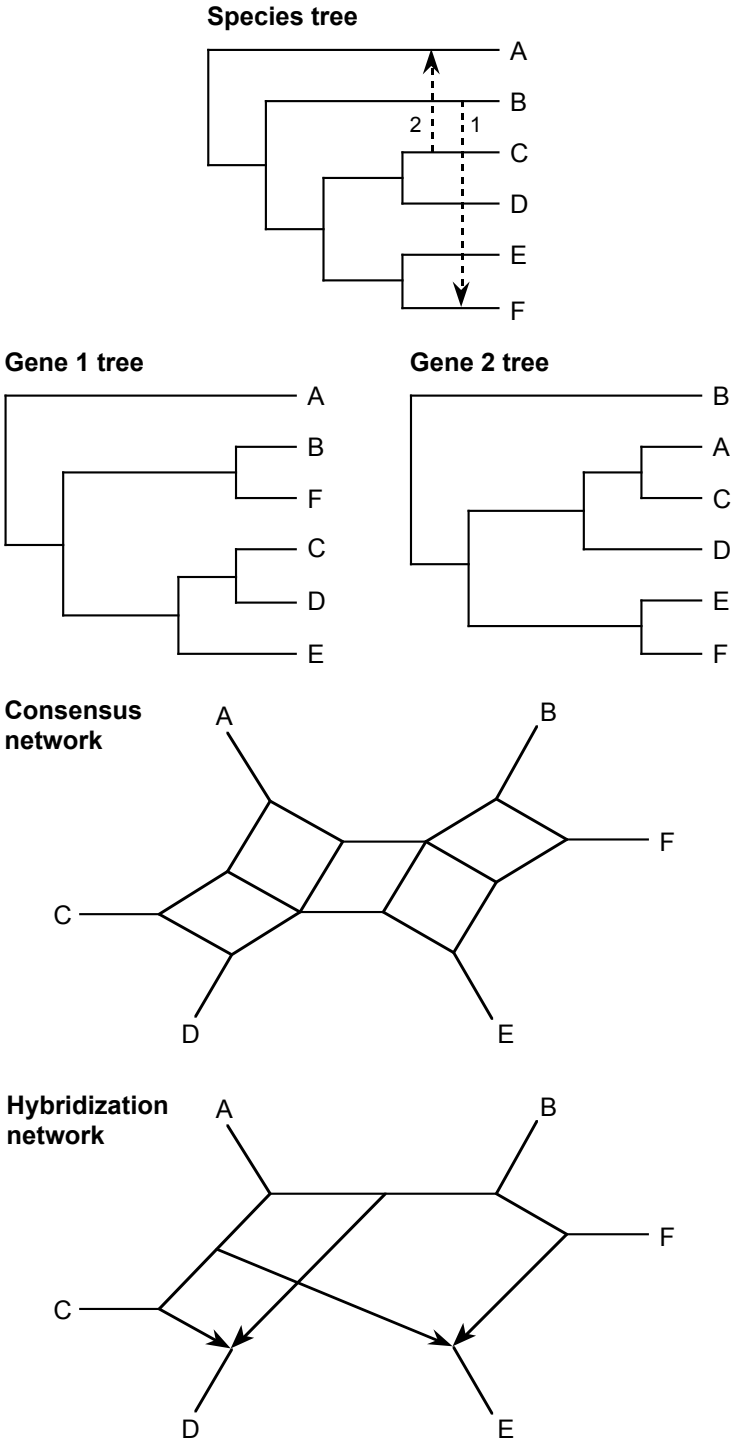

FIG. 14.—Construction of an evolutionary network from a data-display network. The edge lengths

are uninformative. The example species tree (and resulting gene trees) is from Philippe and Douady (2003). The species tree (top) has six species (A–F), and involves two horizontal gene transfers, labelled 1 and 2. Each HGT involves one gene only, producing the two gene trees shown (tree 1 for HGT 1 and tree 2 for HGT 2). The consensus network from these two gene trees is rather uninformative, with five undirected cycles. Resolving this consensus tree into a hybridization network (bottom), rooted on taxon A, produces two tangles (highlighted with arrows), with D and E identified as reticulate taxa. Here, each edge has a direction away from the root, although only four of the edges are shown with arrows.