Consensus Networks: A Method for Visualising Incompatibilities in Collections of Trees

Barbara Holland[1] and Vincent Moulton[2]

[1] Allan Wilson Centre for Molecular Ecology and Evolution, Massey University,
New Zealand.
B.R.Holland@massey.ac.nz
[2] The Linnaeus Centre for Bioinformatics, Uppsala University,
Box 598, 751 24 Uppsala, Sweden.
vincent.moulton@lcb.uu.se

Abstract. We present a method for summarising collections of phylogenetic trees that extends the notion of consensus trees. Each branch in a phylogenetic tree corresponds to a bipartition or split of the set of taxa labelling its leaves. Given a collection of phylogenetic trees, each labelled by the same set of taxa, all those splits that appear in more than a predefined threshold proportion of the trees are displayed using a median network. The complexity of this network is bounded as a function of the threshold proportion. We demonstrate the method for a collection of 5000 trees resulting from a Monte Carlo Markov Chain analysis of 37 mammal mitochondrial genomes, and also for a collection of 80 equally parsimonious trees resulting from a heuristic search on 53 human mitochondrial sequences.

1 Introduction

A central task in evolutionary biology is the construction of phylogenetic trees and, accordingly, many methods have been developed for performing this task. Quite often these methods produce a collection of trees rather than a point estimate of an optimal tree, since such a tree with no measure of reliability may not be particularly helpful. Examples of methods producing collections of trees include Monte Carlo Markov Chain (MCMC) methods [15], [13], and bootstrapping [9]. Heuristic or exact searches [22] can also produce collections of trees if the optimal solution is not unique.

Large collections of trees can be difficult to interpret and draw conclusions from. Thus, when faced with such a collection, it is common practice to construct a consensus tree, i.e., a tree that attempts to reconcile the information contained within all of the trees. Many ways have been devised for constructing consensus trees (see [6] for a comprehensive, recent overview). However, they all suffer from a common limitation: By summarizing all of the given trees by a single output tree, information about conflicting hypotheses is necessarily lost in the final representation.

G. Benson and R. Page (Eds.): WABI 2003, LNBI 2812, pp. 165–176, 2003.

Motivated by this problem we have developed a new approach to visualizing collections of trees that naturally generalizes consensus trees. This approach is based on the construction of phylogenetic networks, networks that are regularly used by biologists to visualize and analyze complex phylogenetic data sets. In particular, we will focus on the use of median networks [3] to visualize collections of trees as we now describe.

2 Methods

First we summarize some necessary concepts.

2.1 Background

Suppose that X is a finite set of taxa. A *split* $A|B$ of X is a bipartition of X, i.e., a partition of X into two non-empty sets or parts A and B with $A \cup B = X$ and $A \cap B = \emptyset$. We call a collection of splits a *split system* for short. A *phylogenetic tree* (on X) is a tree with leaves labelled by X. Each edge of a phylogenetic tree naturally gives rise to a split, since its removal results in two trees, each one being labelled by the elements in one part of a split. We say that a phylogenetic tree *displays* a split if there is an edge in the tree that gives rise to the split. A split system is called *compatible* if there is a phylogenetic tree that displays every split in the system. If this is the case then there is a unique such tree for which the edges are in one-to-one correspondence with the splits in the given system (see e.g., [21, pg 44]). We say that a split system is *incompatible* if it does not contain any subset of cardinality two that is compatible. Note that a split system which is not compatible, need not be incompatible.

It is possible to represent split systems on X by various networks [2], [3], [16]. In particular, a canonical *median network* [4] can be associated to any split system on X. These networks were originally designed for the analysis of mitochondrial data [4] and have also been used to analyze chloroplast data [12]. In a median network, certain vertices are labelled by the elements of X and, in a way similar to phylogenetic trees, splits are represented by classes of parallel edges. Figure 1 illustrates a simple median network on 5 taxa. The median network associated with a split system has several attractive properties. For example, it is a tree if and only if the split system is compatible (in which case it is the unique tree corresponding to the split system), and it is a hypercube if and only if the split system is incompatible [5]. In fact, for a general split system a median network lies somewhere between the extremes of being a tree and a hypercube since each incompatible subsystem of splits with cardinality k corresponds to a k-cube in the network. Moreover, the median network associated with a split system is straight-forward to generate using an algorithm first introduced in [4], which has been implemented in the freely available program Spectronet [11].

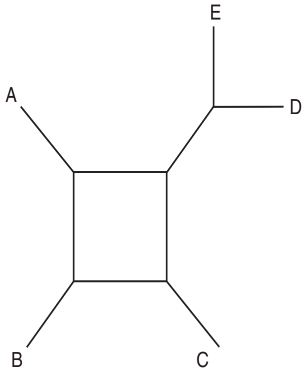


Fig. 1. The median network associated with the split system: $AB|CDE$, $ABC|DE$, $BC|ADE$, $A|BCDE$, $B|ACDE$, $C|ABDE$, $D|ABCE$, $ABCD|E$. The two horizontal parallel edges correspond to the split $AB|CDE$, and the two vertical parallel edges correspond to the split $BC|ADE$.

Since for visual purposes the complexity of the median network associated with a split system is directly related to the degree of incompatibility of the split system (since high dimensional hypercubes are rather difficult to visualize), it is useful to quantify this incompatibility as follows. For k a positive integer, we say that a split system is *k-compatible* if it contains no incompatible subsystem of $k+1$ splits. The concept of k-compatibility was introduced and studied in [8]. Clearly a k-compatible split system is compatible if and only if $k = 1$, in which case its associated median network is a tree, but, as k increases, the associated median network can become progressively more complex. Note that if X has cardinality n, then a (1-)compatible split system on X contains at most $2n - 3$ splits, a 2-compatible split system on the same set contains at most $4n-10$ splits and, for general k, it will contain at most $n(1 + k \log_2(n))$ splits, cf. [8]. Hence for low values of n and k the number of splits in a k-compatible split system on X will not be excessively large, again making the associated median network easier to visualize.

2.2 Consensus Networks

Given a collection of phylogenetic trees, two common methods for computing a consensus tree are the *strict consensus* method, which outputs the tree displaying only those splits that are displayed by all of the input trees, and the *majority-rule consensus* method, which outputs the tree displaying only those splits that are displayed in more than half of the input trees. These two methods can be viewed as being members of a one-parameter family of consensus methods in which a split system S_x is generated that contains precisely those splits that are displayed by more than proportion x of the trees (for strict consensus $x = 1$, and for majority-rule $x = \frac{1}{2}$). If x is greater than $\frac{1}{2}$, then the consensus method results in a split system that is compatible which can thus be displayed by a tree. However, if x is less than $\frac{1}{2}$ this is no longer necessarily the case, although the split system S_x does have the following attractive property.

Theorem 1. *Given N phylogenetic trees and some $0 < x \leq 1$, let S_x denote the split system containing those splits that are displayed in $\lceil Nx \rceil$ or more of these trees. Then S_x is $\lfloor \frac{1}{x} \rfloor$-compatible.*

Proof: Suppose that S_x contains $\lfloor \frac{1}{x} \rfloor + 1$ incompatible splits. Then, since each of these splits is displayed by at least $\lceil Nx \rceil$ of the trees, it follows by the Pigeonhole Principle that one of the trees must display at least two of the incompatible splits. But this is impossible. □

For obvious reasons, we will call the median network associated with S_x a *consensus network*. In order to visualize the contribution that each split makes to the collection of trees in question, we usually weight the edges in this network corresponding to a given split according to the frequency with which it occurs in the trees. This last result indicates a way in which to control the visual complexity of the consensus network associated with S_x. For instance, if we only accept splits that appear in more than $\frac{1}{4}$ of the input trees, then $S_{\frac{1}{4}}$ will be 4-compatible, so that the associated median network is guaranteed to contain cubes only of dimension 3 or less. Note that in the case where $x = 0$, i.e. the split system S contains all splits from all N trees, S is N-compatible.

2.3 Greedy Consensus Networks

We now turn to the practical matter of how to select a split system to be represented by a consensus network. One possibility is to simply select the parameter x described in the previous section by trial and error, and this seems to work reasonably well in practice. A more attractive approach might be to try and select, for fixed k, a maximal k-compatible subset of splits in the split system consisting of all splits displayed by a given collection of trees. However, this is computationally hard even in case $k = 1$ (see e.g., [6]). Even so there are various heuristic approaches possible extending those used to construct consensus trees. We now describe one of these methods.

Consensus trees can be constructed using a *greedy* approach, which can be easily extended to construct networks. We begin by recalling the strategy for constructing a *greedy consensus tree* (cf. [6]). Given a collection of trees, list all splits displayed by at least one of the trees in order of frequency, so that those splits displayed by the largest number of trees come first (with ties broken arbitrarily). A compatible split system is then built up by starting at the beginning of the list and adding in splits one at a time that are compatible with all of the splits in the current split system, ignoring splits that are incompatible with any of the splits in the current system. The tree displaying the resulting compatible split system is the greedy consensus tree.

We construct a *k-greedy consensus network* for a fixed positive integer k in a similar manner, including splits in order of frequency provided they do not lead to a subset of $k + 1$ incompatible splits. As with greedy consensus trees, this approach will also suffer from the fact that if two distinct splits occur with equal frequency, they will be chosen in arbitrary order which can lead to different results (see [6] for more details). In practice we found it useful to stop trying to add further splits after the first split inducing a subset of $k+1$ incompatible splits was obtained; this prevented the main features of the network being obscured by many edges of relatively small weight (results not shown).

2.4 Implementation

Code has been developed to read a list of trees in Newick format (bracket notation) and produce the corresponding weighted split system in nexus format. (Python script available from b.r.holland@massey.ac.nz). This nexus file can then be read by the program Spectronet [11] which displays the associated consensus network.

3 Results

We present two representative examples to illustrate the method.

3.1 MCMC Analysis

Our first example comes from a Monte Carlo Markov Chain (MCMC) analysis [15], [13] of 37 mammal mitochondrial genomes [19]. We used the software MrBayes [14] under a general time-reversible model with gamma distributed rates across sites to generate a chain of 1,000,000 trees; of these every hundredth tree was recorded. We discarded the first half of these trees to provide for a burn in period, leaving 5000 trees in our collection.

Figures 2a-2d show the consensus networks corresponding to the split systems S_x for $x = 1, 0.5, 0.25$ and 0.1. In an MCMC analysis the proportion of times an edge appears in a tree in the chain is interpreted as its posterior probability of being in the true tree, hence the length of the edges in the network are proportional to their posterior probability. Note that all the external edges have posterior probability 1, as they necessarily appear in all of the trees in the collection.

The marsupials (opossum, possum, wallaroo, bandicoot) and the platypus form an outgroup to the placental mammals. We can see in Figures 2a-2d that while the more recent divergences are well resolved, the order of the deeper divergences and the position of the root of the placentals is unresolved. Using the complete data set, the outgroup breaks the rodents into two groups, this is thought to be a long branch attraction artefact, and indeed, when the outgroup taxa are removed the rodents form a single group [19], [20]. Although the strict consensus tree (Figure 2a) and majority-rule consensus tree (Figure 2b) give

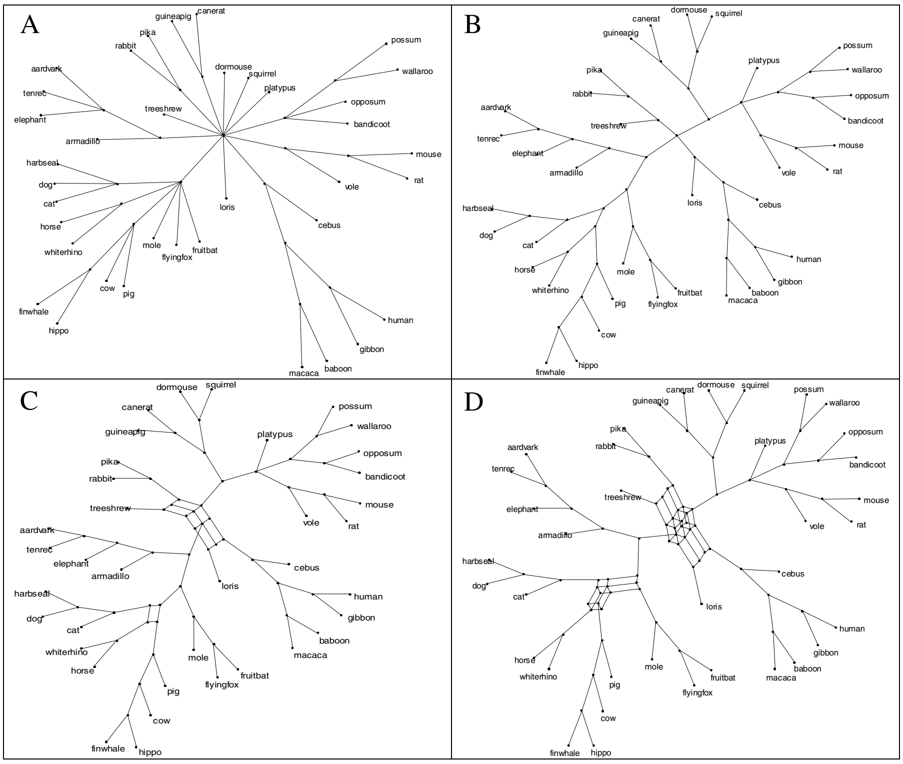


Fig. 2. a) Strict consensus tree for 37 mammal mitochondrial sequences ($x = 1$). b) Majority-rule consensus tree for 37 mammal mitochondrial sequences ($x = 0.5$). c) Consensus network for 37 mammal mitochondrial sequences ($x = 0.25$). This is the smallest value of x for which the associated consensus network contains no 3-cubes. d) Consensus network for 37 mammal mitochondrial sequences ($x = 0.10$). The smallest value of x for which the consensus network contains no 4-cubes is 0.028. However, this network has many tiny edges that detract from the main features.

some idea of the regions of the phylogeny that are uncertain, these regions are displayed either as polytomies, or as edges with weak support, rather than the more informative display of alternative hypotheses in the consensus networks (Figures 2c and 2d). For instance, in Figure 2c there are two possible hypotheses regarding the location of the odd-toed ungulates (horse, white rhino). They could either form a sister group with the carnivores (dog, cat, harbour seal) or with the even-toed ungulates (finwhale, hippo, cow, pig). The relative lengths of edges in the 2-cube indicate that the latter hypothesis is more likely according to this analysis.

3.2 Equally Parsimonious Trees

The second example is a data set consisting of 80 trees. This collection of trees resulted from a heuristic search for the most parsimonious tree for a set of 53 sequences of human mitochondrial DNA [17]. The phylogenetic software package PAUP* [22] was used to search for the maximum parsimony tree, (using the default options Swap=TBR, AddSeq=Simple). All splits appearing in the 80 equally parsimonious trees are shown (Figure 3), this corresponds to $x = 0$, making it unnecessary to compute a greedy network. As we see, rather than sifting through the 80 trees to try and identify similarities and differences, the relevant information is summarized in a single figure.

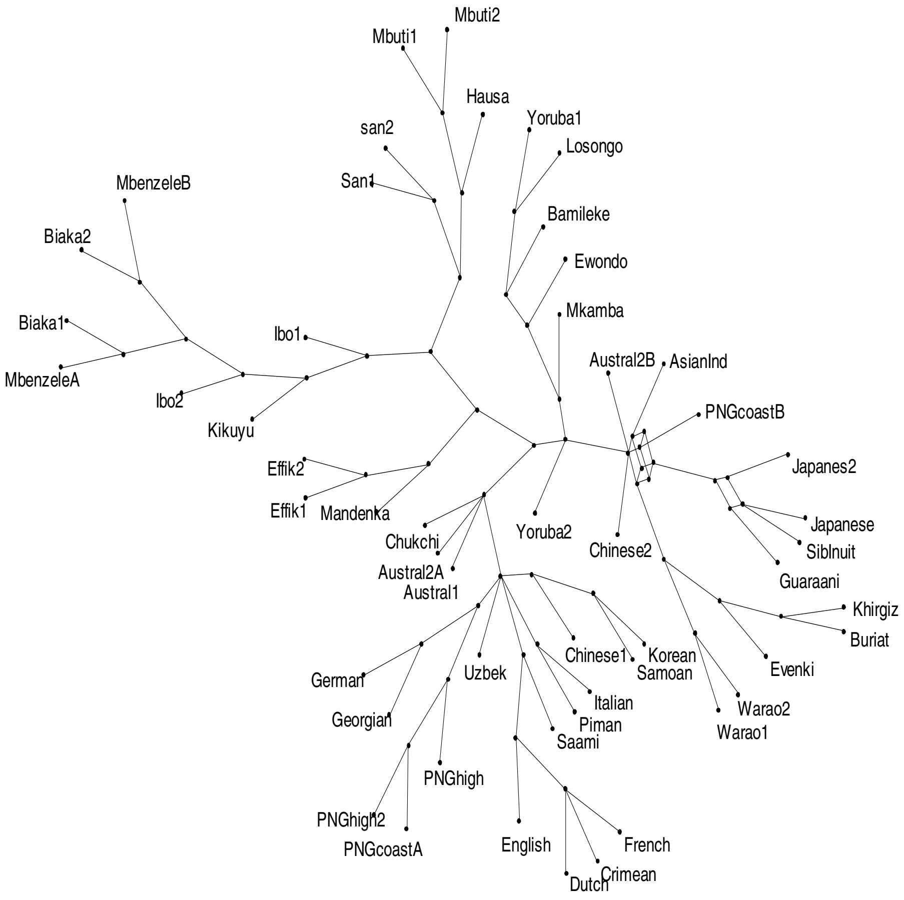


Fig. 3. Consensus network showing all splits in the 80 equally parsimonious trees resulting from a heuristic search on an alignment of 53 human mitochondrial genomes. ($x = 0$).

This data set is typical of intra-species data in that it has many equally likely trees, since taxa are often only separated by a few mutational steps, with a high proportion of the mutations being reversals and parallel changes [4]. As these reversals and parallel changes can lead to conflicting hypotheses about the phylogeny, consensus trees for intra-species data are prone to have many polytomies. This is well illustrated by the majority-rule consensus tree for this data set (Figure 4). There are a large number of resolved trees consistent with the majority-rule tree, only 80 of these are the actual input trees. A greedy consensus tree would provide much greater resolution of the polytomies but would still be unable to display the 3x2 = 6 trees encapsulated by the 3-cube and 2-cube in the network.

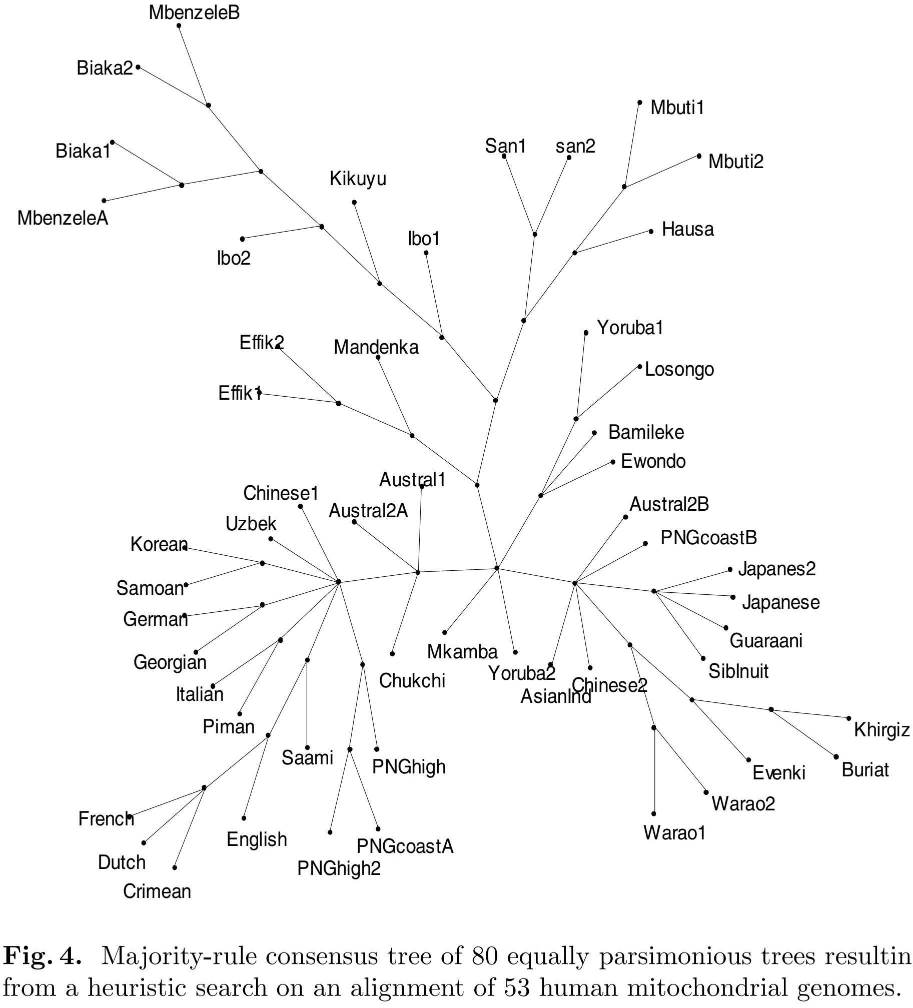


Fig. 4. Majority-rule consensus tree of 80 equally parsimonious trees resulting from a heuristic search on an alignment of 53 human mitochondrial genomes.

The consensus network approach also complements a recent multidimensional scaling method for analyzing collections of trees, TreeViz [18]. This method works by computing a distance between the trees in question (such as the Robinson-Fould's distance), and then using multidimensional scaling to represent the trees as a set of points in a plane (an approach that was also explored in [10]). Using this plot, it is possible to interactively select subcollections of trees and compute consensus trees for these collections.

We show a screenshot of a TreeViz analysis of the 80 equally parsimonious trees (Figure 5a); the multi-dimensional scaling is shown on the left and the consensus tree for 16 selected trees in shown in a panel on the right. In Figure 5b we compare an excerpt from the consensus of the 16 highlighted trees with the corresponding part of the consensus network. Again, where the consensus tree shows a polytomy the network displays the competing hypotheses.

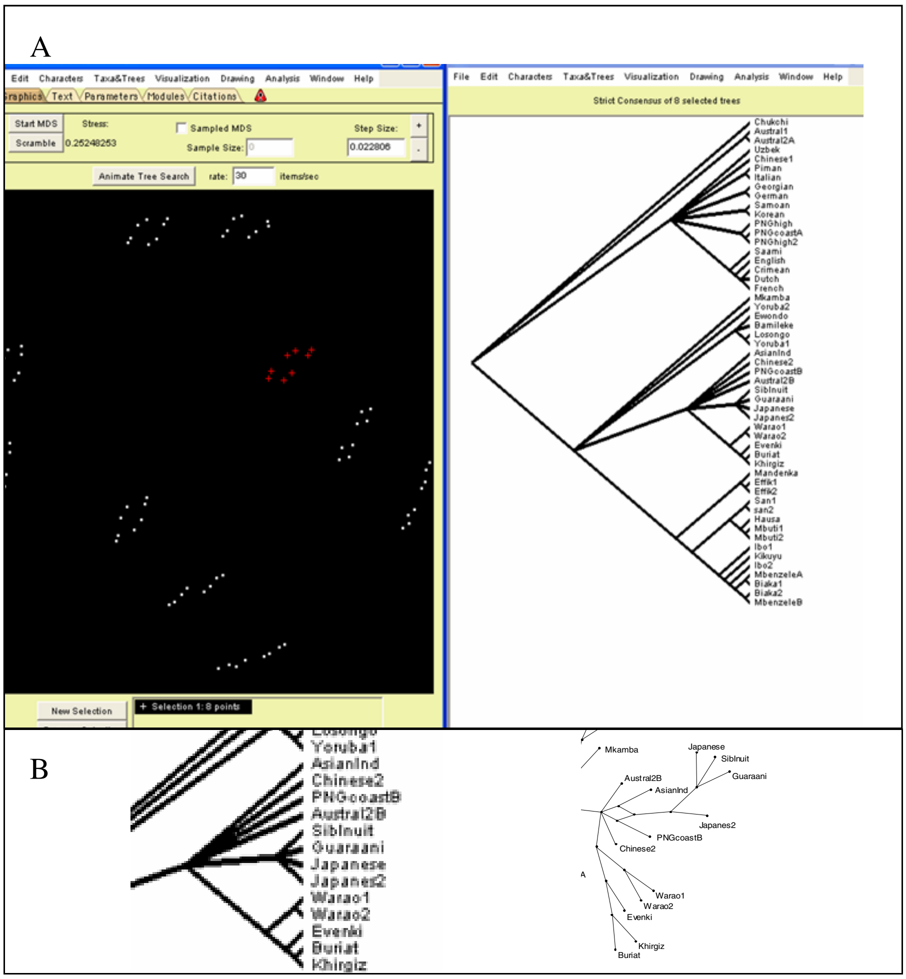
4 Discussion

We have presented a method for generating consensus networks that allows the display of conflicting information within a collection of phylogenetic trees. These networks can be thought of as an extension of strict and majority-rule consensus trees. As with consensus trees, consensus networks can be used as a tool in conjunction with established phylogenetic techniques such as MCMC and bootstrapping. The weights of the edges in consensus networks are open to different interpretations depending on the way in which the input collection of trees is generated. For instance, given a set of trees generated by a MCMC the weights of the splits correspond to posterior probabilities, given bootstrap trees the weights correspond to the confidence level.

One of the main advantages of consensus networks over consensus trees is that they allow conflicting hypotheses within the input collection of trees to be displayed simultaneously in a single diagram. This can be important since a lot of computational effort is usually put into generating large collections of trees, making it somewhat wasteful to only keep a small proportion of this information in the final display. Moreover, it is the conflicts between the trees that are often of interest to biologists and by visual inspection consensus networks allow these to be quickly identified.

Even so, consensus networks still suffer from limitations shared by consensus methods in general. With consensus networks (as with consensus trees) some information may still be lost in order to facilitate display of the network, especially if the data contains many incompatibilities. However, if the data is highly incompatible then it might be questionable in what way a phylogenetic analysis is appropriate.

Another consideration with the consensus networks that we have proposed is that they can still become quite complex, even when restricted to 3-compatible split systems. In practice we found that networks without a distracting level of complexity could be constructed by halting the greedy consensus method



Stress: 0.25248253

□ Sampled MDS

Sample Size: 0

Step Size: 0.022806

rate: 30 items/sec

Selection 1: 8 points

Strict Consensus of 8 selected trees

Fig. 5. a) TreeViz screenshot showing the multi-dimensional scaling of 80 equally parsimonious trees resulting from a heuristic search on an alignment of 53 human mitochondrial genomes (left panel), and the consensus tree for 16 selected trees (right panel). b) A comparison of an excerpt from the consensus of the 16 highlighted trees (5a) with the corresponding part of the consensus network.

after the first split which caused a 4-cube was encountered. Another possible approach to controlling the complexity of the networks is to generate *circular split systems* as opposed to k-compatible split systems. These split systems have the advantage that they can be displayed using split-graphs which, as opposed to median networks, are always planar and can be easily computed using the program SplitsTree [16]. However, we shall explore this possibility elsewhere when we will also look at other avenues for future work including the adaptation of different consensus tree methods to give networks (e.g., matrix representation with parsimony).

References

1. Amenta, N., and Klingner, J.: Case Study: Visualizing Sets of Evolutionary Trees. 8th IEEE Symposium on Information Visualization (InfoVIs 2002)
2. Bandelt, H.-J., and Dress, A.: Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Molecular Phylogenetics and Evolution **1**(3) (1992) 242-252
3. Bandelt, H.-J.: 1994. Phylogenetic Networks. Verhandl. Naturwiss. Vereins Hamburg (NF) **34** (1994) 51–71
4. Bandelt, H.-J., Forster, P., Sykes, B.C., Richards, M.B.: Mitochondrial portraits of human populations using median networks. Genetics **14** (1995) 743–753
5. Bandelt, H.-J., Huber, K.T., Moulton, V.: Quasi-median graphs from sets of partitions. Discrete Applied Mathematics **122** (2002) 23–35
6. Bryant, D.: A classification of consensus methods for phylogenetics. in Janowitz, M.. Lapointe, F.J., McMorris, F., Mirkin, B., Roberts, F. Bioconsensus . DIMACS-AMS. 2003. pp 1–21
7. Bryant, D., Moulton, V.: NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. in the proceedings of WABI, 2002. pp 375–391
8. Dress, A., Klucznik, M., Koolen, J., Moulton, V.: A note on extremal combinatorics of cyclic split systems. Seminaire Lotharingien de Combinatoire **47** (2001) (http://www.mat.univie.ac.at/ slc).
9. Felsenstein, J.: Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39** (1985) 783–791
10. Hendy, M.D., Steel, M.A., Penny, D., Henderson, I.M.: Families of trees and consensus. Pp 355–362, in "Classification and Related Methods of Data Analysis" (H.H. Bock ed.) Elsevier Science Publ. 1988 (North Holland).
11. Huber, K.T., Langton, M., Penny, D., Moulton, V., Hendy, M.: Spectronet: A package for computing spectra and median networks. Applied Bioinformatics **1** (2002) 159–161
12. Huber, K.T., Moulton, V., Lockhart, P., Dress, A.: Pruned median networks: a technique for reducing the complexity of median networks. Molecular Phylogenetics and Evolution **19** (2001) 302–310
13. Huelsenbeck, J. P., Larget, B.,Miller, R.E., Ronquist, F.: Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. **51** (2002) 673-688
14. Huelsenbeck, J. P., Ronquist, F.: MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17** (2001) 754–755
15. Huelsenbeck, J. P., Ronquist, F., Nielsen, R., Bollback, J.P.: Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294** (2001) 2310–2314.
16. Huson, D.: SplitsTree: a program for analyzing and visualizing evolutionary data. Bioinformatics **14** (1998) 68–73 `http://bibiserv.techfak.uni-bielefeld.de/intro/seqdept.html`
17. Ingman, M., Kaessmann, H., Paabo, S., Gyllensten, U.: Mitochondrial genome variation and the origin of modern humans. Science **408** (2000) 708-713

18. Klingner, J.: Visualizing Sets of Evolutionary Trees. The University of Texas at Austin, Department of Computer Sciences. Technical Report CS-TR-01-26. (2001)
19. Lin, Y.-H., McLenachan, P.A., Gore, A.R., Phillips, M.J., Ota, R., Hendy, M.D., Penny, D.: Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Molecular Biology and Evolution **19** (2002) 2060–2070
20. Lin, Y.-H., Waddell, P.J., Penny, D.: Pika and Vole mitochondrial genomes add support to both rodent monophyly and glires. Gene **294** (2002) 119-129
21. Semple, C., Steel, M.: Phylogenetics, Oxford University Press 2003.
22. Swofford, D.L.: PAUP* - Phylogenetic Analysis Using Parsimony (*and other methods) Version 4. Sinauer Associates, Sunderland, Mass. 1998.