Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks

David Bryant and Vincent Moulton†*

*McGill Centre for Bioinformatics, Montreal, Quebec, Canada; †The Linnaeus Center for Bioinformatics, Uppsala University, Uppsala, Sweden

We present Neighbor-Net, a distance based method for constructing phylogenetic networks that is based on the Neighbor-Joining (NJ) algorithm of Saitou and Nei. Neighbor-Net provides a snapshot of the data that can guide more detailed analysis. Unlike split decomposition, Neighbor-Net scales well and can quickly produce detailed and informative networks for several hundred taxa. We illustrate the method by reanalyzing three published data sets: a collection of 110 highly recombinant Salmonella multi-locus sequence typing sequences, the 135 ''African Eve'' human mitochondrial sequences published by Vigilant et al., and a collection of 12 Archeal chaperonin sequences demonstrating strong evidence for gene conversion. Neighbor-Net is available as part of the SplitsTree4 software package.

Introduction

The Neighbor-Joining (NJ) method of Saitou and Nei (1987) is arguably the most widely used distance-based method for phylogenetic analysis. The NJ algorithm takes an arbitrary distance matrix and, using an agglomerative process, constructs a fully resolved (bifurcating) phylogenetic tree. Dozens of simulation studies and thousands of phylogenetic analyses have demonstrated that NJ is both fast and quite accurate. This success has in turn inspired many variations on the theme, including BioNJ and UNJ (Gascuel 1997*a*, 1997*b*), Weighbor-Joining (Bruno, Socci, and Halpern 2000), and NJML (Ota and Li 2000). In this article we describe a new variant of NJ, one which constructs phylogenetic networks instead of phylogenetic trees.

Phylogenetic networks generalize phylogenetic trees because they permit the representation of conflicting signal or alternative phylogenetic histories (Fitch 1997). The use of networks, rather than simple branching trees, is clearly necessary when the underlying evolutionary history is not treelike. Recombination, hybridization, gene conversion, and gene transfer all lead to histories that are not adequately modelled by a single tree. Even when the underlying history *is* treelike, parallel evolution, model heterogeneity, and sampling error may make it difficult to determine a unique tree. In these cases networks can provide a valuable tool for representing ambiguity or for visualizing a space of feasible trees.

There are a number of phylogenetic network methods already in use—Posada and Crandall (2001) provide a comprehensive review. The methods divide roughly into two classes. The first class includes methods that construct networks directly from character data, typically under a parsimony framework. The nodes in the network represent taxa (for example, different haplotypes), hypothetical ancestral taxa, or intermediary nodes. The best-known network methods in this class are *statistical parsimony* (Templeton, Crandall, and Sing 1992), *median networks* (Bandelt et al. 1995), the variants of median networks (Bandelt, Forster, and Röhl 1999; Huber et al. 2001, 2002), and the *netting* method (Fitch 1997). These methods are designed for the analysis of intraspecific data. They often run into problems when the level of diversity increases, either because the networks become too complicated, or through the increasing influence of reduction rules on the resulting network. As well, increased diversity can lead to the same consistency problems encountered with parsimony unless hidden and parallel mutations are corrected for (as in SpectroNet [Huber et al. 2002]).

The second major class of phylogenetic network methods includes those that construct networks directly from a distance matrix. The use of distance data alone means that these phylogenetic network methods start with less information than those using the complete alignment. Nevertheless, there is evidence that a lot of phylogenetic information is preserved in the distance matrix, even in the presence of reticulation (Bryant et al. 2003; Legendre and Makarenkov 2002; Xu 2000).

Neighbor-Net is a distance-based method. It is most closely related to Pyramid Clustering and Split Decomposition. Pyramid clustering, like Neighbor-Net, works agglomeratively (Diday 1986). The relationship between Pyramid clustering and Neighbor-Net is loosely analogous to that between UPGMA and NJ, although the agglomeration and reduction processes used in the two methods are quite different. Split decomposition (Bandelt and Dress 1992), implemented in SplitsTree (Huson 1998), decomposes the distance matrix into simple components based on weighted *splits* (bipartitions of the taxa set). These splits are then represented using a *splits graph*, a special type of phylogenetic network that simultaneously represents both groupings in the data and evolutionary distances between taxa (see later under *Splits Graphs*). Neighbor-Net works in a similar way: we first construct a collection of weighted splits, then represent these splits using a splits graph. The advantage of Neighbor-Net is that it tends to construct networks that are much more resolved than those given by split decomposition.

Methods

Background—Compatible and Incompatible Splits

A *split* is a partition of the set of taxa into two disjoint, non-empty groups. Splits are the building blocks

Key words: networks, Neighbor-Joining, recombination, SplitsTree.

E-mail: bryant@mcb.mcgill.ca.

Mol. Biol. Evol. 21(2):255–265. 2004
DOI: 10.1093/molbev/msh018
Advance Access publication December 5, 2003
Molecular Biology and Evolution vol. 21 no. 2

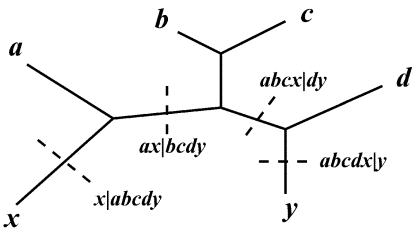

FIG. 1.—The splits graph for a set of compatible splits (i.e., a tree). The splits for the branches along the path from x to y are given. These are exactly the splits of the tree with x and y in different groups.

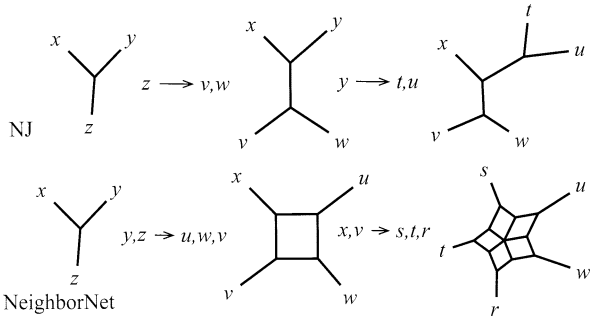

FIG. 3.—Expanding after agglomeration: NJ and Neighbor-Net.

of phylogenetic trees: each branch divides the set of taxa up into a split, with the taxa on one side of the branch separated from the taxa on the other side (fig. 1). The collection of splits given by all the different branches in an unrooted phylogeny T contains all of the branching information of the phylogeny. We call this collection the *set of splits* of T.

A collection of splits is *compatible* if it is contained within the set of splits of some phylogenetic tree; otherwise it is *incompatible*. When we construct phylogenies we construct compatible collections of splits. To generalize trees, we must allow collections of splits that are incompatible. As we shall see, the collections constructed by Neighbor-Net are not, in general, compatible, but instead satisfy a weaker condition than compatibility.

We will be dealing with *weighted* collections of splits. The weights for a compatible collection of splits correspond to the lengths of the corresponding branches. Recall that the distance between any two taxa x, y in a tree, also called the *phyletic distance* (Fitch 1997), equals the sum of the lengths of the branches along the path from x to y. The branches along this path correspond exactly to the splits in the tree that have x and y on opposite sides (fig. 1). Hence the phyletic distance between x and y equals the sum of the split weights for all those splits having x and y in different groups.

This formulation of phyletic distance extends directly to collections of splits that are not compatible. The phyletic distance between two taxa, with respect to a collection of weighted splits, equals the sum of the weights of the splits separating them. This in turn equals the length of a shortest path between the two taxa in the splits graph representation of the collection of splits (see later under *Splits Graphs*).

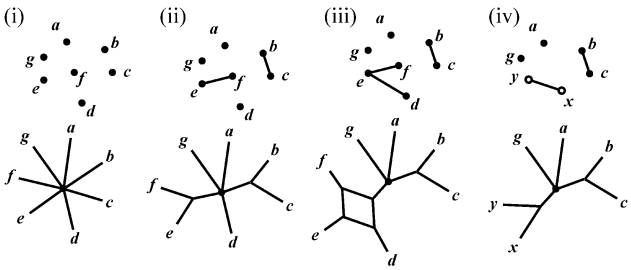

FIG. 2.—The agglomerative process for Neighbor-Net. (i) We begin with each node representing a single taxon. (ii) Using the selection criterion, we identify b and c as neighbors, as well as e and f. Unlike NJ, we do not amalgamate immediately. (iii) We have identified e as a neighbor of d (as well as f). Notice how the splits $ef|abcdg$ and $de|acdfg$ are both represented in the splits graph. (iv) As e has two neighbors, we perform a reduction, replacing d, e, f by x, y.

The Neighbor-Net Method

The Agglomerative Process

NJ, UPGMA, and the linkage tree algorithms all follow the same general scheme. We start with one node for each taxon. At each iteration, a pair of nodes is selected and replaced by a new composite node. The procedure continues until only two (or three) nodes remain, at which point we reverse the process to construct a tree or hierarchy.

Neighbor-Net works along the same lines, with one important difference. When we select a pair of nodes we do not combine and replace them immediately. Instead we wait until a node has been paired up a second time. We replace the three linked nodes with two linked nodes and reduce the distance matrix. If there is still a node linked to two others, we perform a second agglomeration and reduction. We then proceed to the next iteration. This simple change in the agglomerative framework generates a collection of splits that cannot be represented by a single tree. The process is illustrated in figure 2.

With NJ, we amalgamate pairs of nodes into a single node, repeating the process until only three nodes remain. If we keep a list of these amalgamations, we can reconstruct the NJ tree by reversing the amalgamation process (fig. 3). With Neighbor-Net we also keep a list of amalgamations, though each amalgamation replaces three nodes with two. Reversing the amalgamation process gives the splits in the Neighbor-Net (fig. 3).

The end-product of the Neighbor-Net process is a *circular* collection of splits, as can be proved using mathematical induction. Circular collections of splits are a mathematical generalization of compatible collections of splits. Formally, a collection of splits of X is circular if there is an ordering $x_1, x_2, \ldots, x_n$ of the taxa such that every split is of the form $\{x_i, x_{i+1}, \ldots, x_j\} \,|\, X - \{x_i, \ldots, x_j\}$ for some i and j satisfying $1 \leq i \leq j < n$. Graphically, circular splits arise when we place the taxa around a circle and consider the splits given by cutting the circle along a line (fig. 4). Most importantly, Andreas Dress and Daniel Huson (personal communication) have proved that circular

collections of splits always have a planar splits graph representation (see later under *Splits Graphs*).

Within the agglomerative framework, the Neighbor-Net method is determined by the formulae used to select nodes for agglomeration and the formula used to reduce the distance matrix after each agglomeration.

Selection Formulae

The selection formulae are closely related to the formulae used for NJ. Suppose that we have n nodes remaining. At the very beginning of the algorithm, none of the nodes will have neighbors already assigned to them. Later on, some pairs of nodes will have been identified as neighbors, but not agglomerated. We need to take these neighbor relations into account when selecting nodes to agglomerate.

The neighboring relations group the n nodes into clusters $C_1, C_2, \ldots, C_m$, $m \leq n$, some of which contain a single node and others of which contain a pair of neighboring nodes. The distance $d(C_i, C_j)$ between two clusters is the average of the distances between elements in each cluster:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d_{xy}. \tag{1}$$

The selection of neighboring nodes proceeds in two steps. First we find the pair of clusters that minimizes the standard NJ formula

$$Q(C_i, C_j) = (m-2)d(C_i, C_j) - \sum_{\substack{k=1 \\ k \neq i}}^{m} d(C_i, C_k) - \sum_{\substack{k=1 \\ k \neq j}}^{m} d(C_j, C_k). \tag{2}$$

Suppose that C_{i^*} and C_{j^*} are two clusters that minimize $Q(C_i, C_j)$. The second step is to choose which node $x_i \in C_{i^*}$ and which node $x_j \in C_{j^*}$ are to be made neighbors. The clusters C_{i^*} and C_{j^*} each contain either one or two nodes. If these clusters were separated into individual nodes we would end up with $m + |C_{i^*}| + |C_{j^*}| - 2$ clusters in total. Let $\hat{m}$ denote $m + |C_{i^*}| + |C_{j^*}| - 2$. To maintain consistency, this value $\hat{m}$ replaces m in equation (2) when we are selecting particular nodes *within* clusters. That is, we select the node $x_i \in C_{i^*}$ and node $x_j \in C_{j^*}$ that minimizes

$$\hat{Q}(x_i, x_j) = (\hat{m}-2)d(x_i, x_j) - \sum_{\substack{k=1 \\ k \neq i}}^{\hat{m}} d(x_i, C_k) - \sum_{\substack{k=1 \\ k \neq j}}^{\hat{m}} d(x_j, C_k). \tag{3}$$

The choice of selection formulae, and the reduction formula which follows, guarantees the statistical consistency of the Neighbor-Net method. We discuss consistency below.

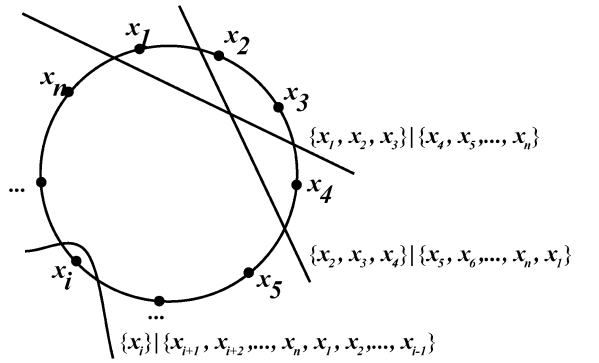

FIG. 4.—A collection of splits is circular if it corresponds to cuts of a circle.

Distance-Reduction Formulae

Suppose that node y has two neighbors, x and z. In the Neighbor-Net agglomeration step, we replace x, y, z with two new nodes u, v.

The distances from u and v to another node a are computed using the reduction formulae

$$\begin{aligned} d(u,a) &= \alpha\, d(x,a) + \beta\, d(y,a) \\ d(v,a) &= \beta\, d(y,a) + \gamma\, d(z,a) \\ d(u,v) &= \alpha\, d(x,y) + \beta\, d(x,z) + \gamma\, d(y,z) \end{aligned}$$

where α, β, γ are non-negative real numbers with $\alpha + \beta + \gamma = 1$.

Gascuel (1997a) observed that a single degree of freedom can be introduced into the reduction formulae for NJ. In the above formulae we introduce two degrees of freedom, thereby opening the possibility for a variance reduction method in future versions of Neighbor-Net. By default we use $\alpha = \beta = \gamma = \frac{1}{3}$, the equal coefficients being directly analogous to NJ.

Estimating Split Weights

The NJ algorithm computes both a tree and branch lengths for that tree. The branch lengths are computed while the tree is being constructed, using a variant of the least squares formulae. We also use the least squares framework for Neighbor-Net.

As we observed above, the phyletic distance between two taxa equals the sum of the weights of the splits that separate them. Suppose that the splits in the network are numbered 1, 2, ..., m and that the taxa are numbered 1, 2, ..., n. Let $\mathbf{A}$ be the $n(n-1)/2 \times m$ matrix with rows indexed by pairs of taxa, columns indexed by splits, and entry $\mathbf{A}_{(ij)k}$ given by

$$\mathbf{A}_{(ij)k} = \begin{cases} 1 & \text{if } i, j \text{ are on opposite sides of split } k; \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The matrix $\mathbf{A}$ is the network equivalent of the standard topological matrix for a tree (see, e.g., Cavalli-Sforza and Edwards [1967]; Farris [1972]). The matrix corresponding to the network in figure 5 equals:

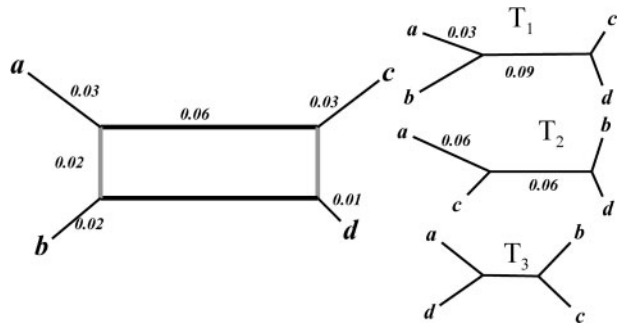

FIG. 5.—The simplest splits graph that is not a tree. The graph "contains" T_1 and T_2 but not T_3.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}. \quad (5)$$

We represent a distance matrix d by an $n(n-1)/2$ dimensional vector

$$\mathbf{d} = (d_{12}, d_{13}, \ldots, d_{(n-1)n})'. \quad (6)$$

The corresponding vector of phyletic distances is given by $\mathbf{p} = \mathbf{Ab}$.

Neighbor-Net produces a circular collection of splits, so the corresponding matrix $\mathbf{A}$ has full rank (Bandelt and Dress 1992). The ordinary least squares (OLS) estimates for $\mathbf{b}$ can therefore be computed from the observed distance vector $\mathbf{d}$ using the standard formula

$$\mathbf{b} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{d}. \quad (7)$$

Weighted least squares (WLS) estimates can be computed using

$$\mathbf{b} = (\mathbf{A}'\mathbf{W}\mathbf{A})^{-1}\mathbf{A}'\mathbf{W}\mathbf{d} \quad (8)$$

where $\mathbf{W}$ is the $n(n-1)/2 \times n(n-1)/2$ diagonal matrix with $1/\mathrm{var}(d_{ij})$ in entry $\mathbf{W}_{(ij)(ij)}$. These formulae are identical to those used for phylogenies (Cavalli-Sforza and Edwards 1967; Farris 1972).

Equations (7) and (8) can produce some negative split weights. If we discard splits with negative weight and leave the remaining splits unchanged, the weights of the remaining splits are often grossly overestimated. The positively weighted splits compensate for the negatively weighted splits. Consequently, many more redundant splits are retained, creating a network that is both overly complicated and a poor fit to the data. The same problem can hold for phylogenies, but the situation for networks is more serious.

For this reason, we always compute optimal least squares estimates with a non-negativity constraint. There is no closed formula for constrained least squares estimates. Enforcing the constraint increases computation time considerably, but the result is a far cleaner and more accurate representation.

Splits Graphs

Neighbor-Net constructs a collection of weighted splits which is then converted to a graphical representation, called a *splits graph*, using the drawing algorithms implemented in SplitsTree (Huson 1998). A splits graph is a graphical representation of a collection of weighted splits. The splits graph for a compatible collection of splits is precisely a tree: each edge in the graph corresponds to a split in the collection and has length equal to the weight of the split. Incompatible splits are represented by splits graphs with cycles or boxes. Each split in a splits graph then corresponds to a *collection* of parallel edges, all with the same length. Removing the edges corresponding to a given split $A \mid B$ partitions the network into two connected parts, one containing the taxa in A and the other containing the taxa in B.

In a tree, the phyletic distance between two taxa equals the sum of the lengths of the path connecting them. The presence of cycles in a splits graph means that there can be several paths between any two taxa. The phyletic distance between two taxa x, y equals the length of a shortest path connecting them. One can show (A. Dress and D. Huson, personal communication) that these shortest paths include exactly one edge corresponding to every split in the graph separating x from y. Hence the distance between x and y also equals the sum of the split weights of those splits separating x and y.

Some examples: The simplest splits graph that is not a tree is depicted in figure 5. The graph represents six splits: the four splits separating one taxa from the rest, one split separating a, b from c, d, and another separating a, c from b, d. The two darker internal edges correspond to the split $\{a, b\} \mid \{c, d\}$, and the gray edges correspond to $\{a, c\} \mid \{b, d\}$. Split weights are marked on the graph.

The interpretation of these graphs depends on the significance of the corresponding splits and their weights. Both trees T_1 and T_2 have their splits contained in the splits graph, but T_3 does not. If the splits graph is taken to represent a distance matrix between a, b, c, d, we can see that this distance matrix is closer to the distance matrix given by T_2 than T_1. The splits graph can also represent mixtures of two trees. The weights in the example are consistent with a mosaic alignment where $\frac{2}{3}$ of the sites support T_1 and $\frac{1}{3}$ support T_2. The weight for $\{a, b\} \mid \{c, d\}$ in the splits graph (0.06) equals the weight of the split in T_1 (0.09) multiplied by the proportion (2/3) of sites supporting that tree. The split $\{a\} \mid \{b, c, d\}$ appears in T_1 with weight 0.03 and in T_2 with weight 0.06. The weight in the splits graph is therefore $2/3 \times 0.03 + 1/3 \times 0.06 = 0.03$.

The splits graph (i) in figure 6 is more complicated. Three pairwise incompatible splits generate a three-dimensional, non-planar, cube. However, this splits graph can be simplified: the splits graph in figure 6 (ii) displays exactly the same splits. The information represented by both networks is identical.

This example illustrates two points: the splits graph representation need not be planar and it is not necessarily unique. This first problem is not an issue for us: the splits graphs generated by Neighbor-Net are *always* planar, an important advantage over other network methods when it

comes to visualization. The second problem means that care must be taken when interpreting internal, or ancestral, nodes in the graph. A splits graph represents conflict, and conflicting signals, rather than an explicit history of which reticulations took place (Strimmer, Wiuf, and Moulton 2001). That said, boxes in the splits graph can be used to locate reticulations which can then be validated by other techniques.

Consistency

Neighbor-Joining is consistent. If the input to NJ is a distance matrix that is already additive (treelike), then NJ will return the corresponding weighted phylogenetic tree (see (Gascuel 1997*a*) for a review). This condition guarantees statistical consistency under a wide range of stochastic models.

Neighbor-Net is also consistent. If the input to Neighbor-Net is a treelike distance matrix, Neighbor-Net will return the splits and branch lengths of the corresponding tree. In fact, Neighbor-Net is consistent for all circular distance matrices, a much wider class of distance matrices. A distance matrix is *circular* (also called *Kalmanson*) if it equals the phyletic distances for a circular collection of splits with positive weights. Because compatible splits are circular, treelike (or *additive*) distances are circular. If the input distance matrix is circular, Neighbor-Net is guaranteed to return the corresponding circular splits with their split weights. The proof is non-trivial—refer to Bryant and Moulton (2003) for details. This consistency property explains (and, in fact, almost determines) the specific choice of selection and reduction formulae presented above.

Examples

To illustrate the application of Neighbor-Net, we reanalyzed three published data sets using Neighbor-Net. The distance matrices used, and examples of further studies and simulations, are available online from the Neighbor-Net Web page (http://www.mcb.mcgill.ca/~bryant/NeighborNet). Neighbor-Net itself is available as part of the SplitsTree 4.0 software package.

Salmonella MLST Data

Kotetishvili et al. (2002) describe the use of multilocus sequence typing (MLST) to classify several hundred *Salmonella* isolates. Split decomposition was used to test for the presence of recombination within the data set. The authors detected evidence for recombination in two of the genes studied, but they were forced to reduce the number of taxa for the analysis of the phosphomannomutase (*manB*) sequences. We therefore repeated the analysis of all the 110 *manB* sequences using Neighbor-Net.

We first estimated evolutionary distances using maximum likelihood, with parameters determined using Modeltest (Posada and Crandall 1998). The network produced by Neighbor-Net permitted the selection of a small group of sequences that were subsequently tested for recombination using the LikeWin software (Archibald and Roger 2002*a*).

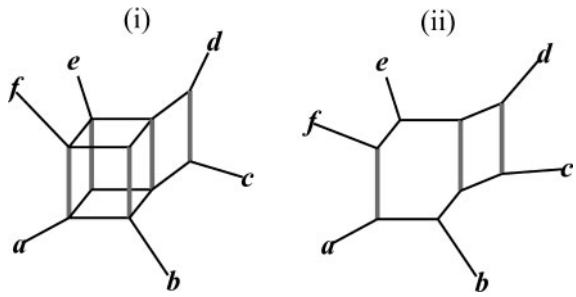

FIG. 6.—Two splits graph representations of the same set of splits. The edges corresponding to the split $\{a, b, c\} \mid \{d, e, f\}$ are highlighted in both graphs.

LikeWin uses PAUP* (Swofford 1998) to compute a maximum likelihood tree T for all of the sites. A sliding window of width 100 is then moved along the sequence. For each window we compare the maximum likelihood score for a tree on those sites to the likelihood of T. A significant difference between these two indicates a change in signal. Statistical significance was estimated using parametric bootstrapping, repeating the entire sliding window analysis on multiple (we used 100) simulated sets of sequences (following Archibald and Roger [2002*a*]).

Mitochondrial Eve Data

Our second example revisits the phylogenetic analysis of 135 human mitochondrial sequences, originally published by Vigilant et al. (1991). A phylogeny for these sequences was used as supporting evidence for an African origin of human mitochondria. The validity of this study was later questioned, though an extensive study of the large-scale landscape of the space of trees (Penny et al. 1995) indicates that data support the phylogenetic hypotheses put forward by Vigilant et al. (1991).

The central problem with these data is the large number of sequences and the small number of sites. Sampling error leads to substantial homoplasy between the sequences, and the relative lack of information in the data means that there will be millions, perhaps billions, of optimal parsimony trees. This is an ideal situation for a network analysis, because we can deduce features from the data without restricting our attention to a single tree.

We estimated distances from the mitochondrial sequences using K2P + Γ (=0.5). Following Penny et al. (1995), we reweighted the characters to compensate for hypervariable sites (site weights kindly supplied by D. Penny) and constructed the Neighbor-Net.

Archeal Chaperonin Data—Gene Conversion

For the third example, we reanalyzed DNA sequence data from the chaperonin complexes of 12 crenarchaeotes (Archea), originally published by Archibald and Roger (2002*b*). The taxa divide into α and β paralogs stemming from an ancestral duplication. Archibald and Roger find substantial evidence of gene conversion between the two paralogs. Indeed, the presence of some gene conversion between different paralogs is obvious from a visual inspection of the alignments, most significantly with

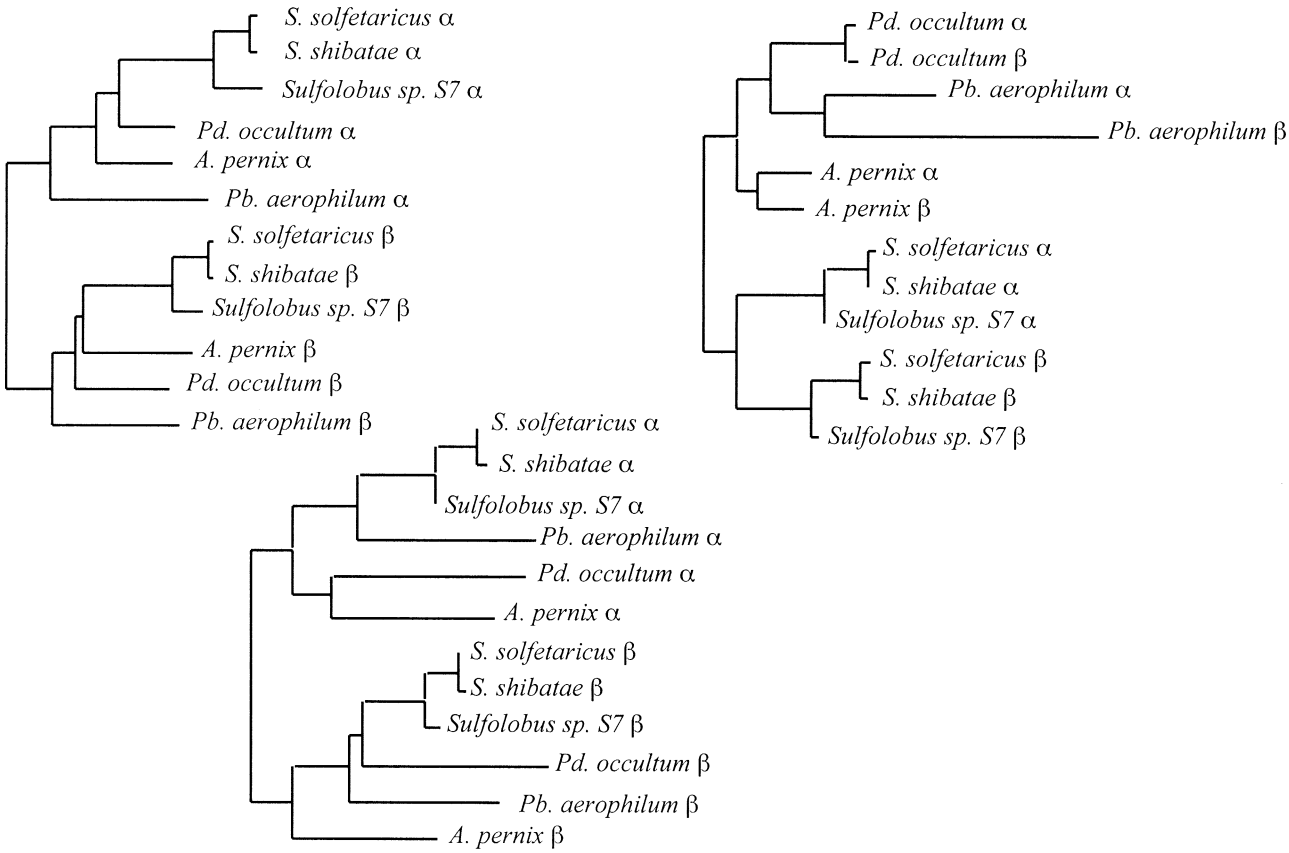

FIG. 7.—Maximum likelihood trees for three regions within the Archael Chaperonin sequences. (Adapted from Archibald and Roger [2002*b*].)

Pyrodictium occultum. Despite this, Geneconv (Sawyer 1989) gave only weak or moderate support for the presence of gene conversion, especially once selection pressures had been corrected for (Archibald and Roger 2002*b*). Archibald and Roger also report that the likelihood-based software of Grassly and Holmes (1997) and the distance-based software of McGuire and Wright (2000) both failed to detect significant gene conversion. They therefore developed a new sliding-window method, validated by parametric bootstrapping, that was used to identify three principal domains supporting the three phylogenies presented in figure 7.

We computed a distance matrix directly from the alignment using ML distances (parameters taken from Archibald and Roger [2002*b*]) and performed a Neighbor-Net analysis.

Results

Salmonella Data

The Neighbor-Net for the 110 MLST *manB* sequences is given in figure 8. Whereas split decomposition returns a tree for the full data set (Kotetishvili et al. 2002), the Neighbor-Net is distinctly non-treelike. However the presence of boxes in the network does not imply recombination, only the possibility of recombination (see the human mitochondrial analysis below). Therefore we applied the sliding-window technique developed by Archibald and Roger (2002a) to test for recombination.

The sliding-window analysis, and particularly the parametric bootstrap method used to test significance, requires a huge amount of computation, and is infeasible for the complete set of 110 sequences. Instead we used Neighbor-Net to select a small set of taxa to test for recombination in a specific area in the network (identified in figure 8). We used the same model parameters as those determined for the complete set of sequences.

The analysis detected two areas where the window likelihoods differed substantially from that for the complete set of sites (fig. 9). The significance levels, estimated using a parameterized bootstrap, are ($P < 0.03$) for the larger peak and ($P < 0.23$) for the lower peak. Note that the significance test is based on a single window maximum and says nothing on the significance of observing the multiple adjacent windows with high likelihood differences that we encounter here. We focused on the large peak, estimated breakpoints roughly from the LikeWin graph, and repeated the Neighbor-Net analysis including and excluding different sites (fig. 10).

Our first observation is that removing sites 110–250 removes almost all boxes from the restricted network. Thus, in these seven sequences, we can conclude that much of the conflicting signal comes from these sites. The second and perhaps most important observation is that the divergence between the taxa within sites 110–250 is

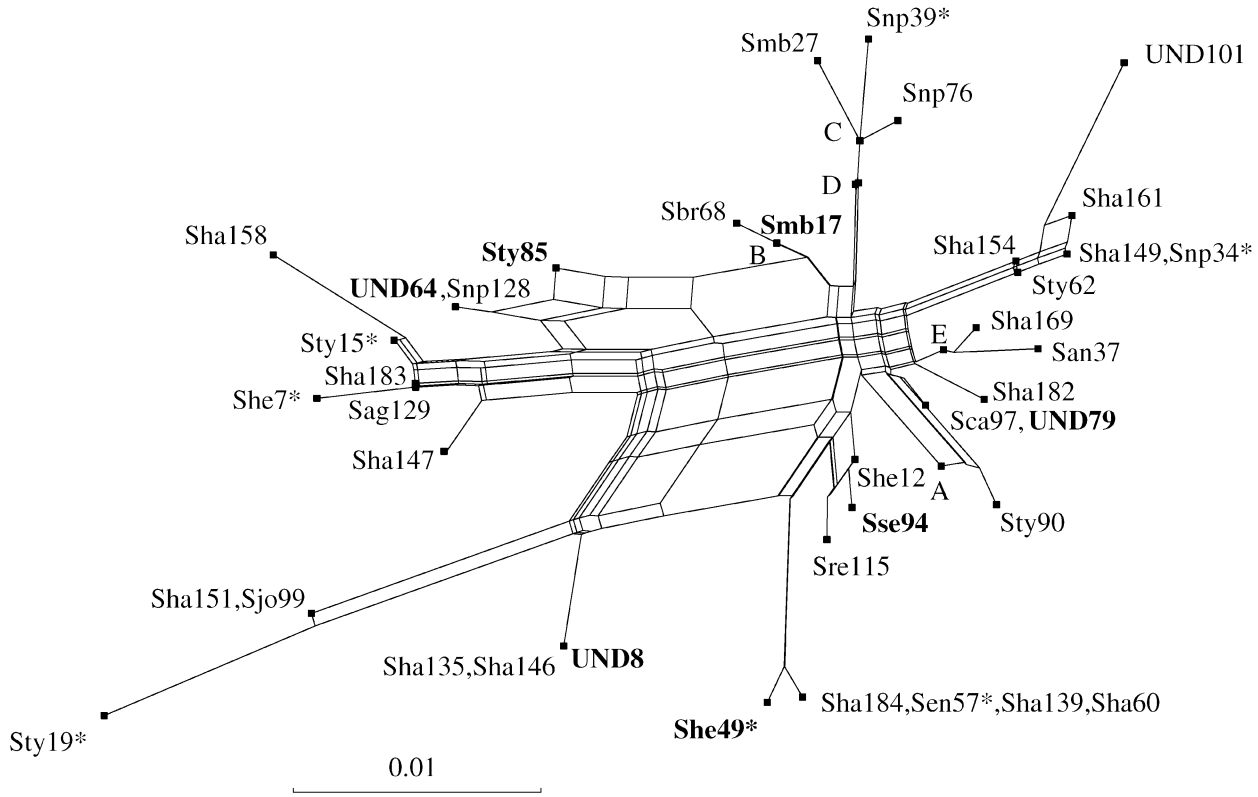

FIG. 8.—Neighbor-Net for the 110 Salmonella *manB* sequences. The isolates used for the sliding-window analysis are in boldface. Clinical isolates indicated by an asterisk (*). Group A includes the isolates Sty54, Sty54*, Sty2, She9, Sty87, Snp40*, Sty13, Snp41*, Sen5, Sha160, Sha141, Sty20*, Sha58, Sse18, Sha71, Sty31. Group B includes the isolates Sty61, Sha148, Smb-17, Sag75, Sha124. Group C includes the isolates UND3, Sha150, Sha173, Sen23*, Sha153, Sha140, San96, Sen30*, Sen24*, Sha138, Sha176, Sha130, Sha164, Sha157, Sen29*, Sca93, Sha122, Sht20, Sha186. Group D includes the isolates She3, Sha50, Sse95, Sha56, Sen24, Sen34, Sha177, Sty13*, Swo44, Sty86, Ste41, Sha77, UND80. Group E includes the isolates Ssc40, Sse28, Sty89, Sty15*, Ske69, UND110, Sha49, Sen4, Sha48, Sha165, Sty92, Snp33*, Sty52, UND109, Sha131, Sha102, Sty6, Sha175.

significantly higher than for the remaining sites. This suggests that a change in rate could have made a major contribution to the difference in log likelihoods found using the sliding-window analysis. In contrast, an investigation of the second smaller peak (sites 340–450) revealed an area where divergence was significantly lower than in the remaining sites.

We repeated the partitioned analysis for all 110 sequences (networks not shown). The network for sites 110–250 was almost identical to that for the seven sequences, with one major conflicting split. The network for the remaining sites was not treelike, but it had significantly fewer boxes than the network for all of the sites. Finally, the network for sites 340–450 was completely treelike, dividing the sequences into only six groups.

Clearly, further analysis is required to unravel the evolutionary history of these sequences. Our aim here is to illustrate how Neighbor-Net might be used to guide more detailed investigations.

Mitochondrial Eve Data

The Neighbor-Net for the 135 human mitochondrial sequences indicates very clearly why these data have been so difficult to analyze (fig. 11). The network represents marked ambiguity in the signal. There are conflicting splits throughout the network. Even areas that appear treelike (such as the area around the Asian–European groups) are in fact full of boxes, as can be most clearly demonstrated by manipulating the network within SplitsTree.

Given the history and context of the data, it is reasonable to infer that the ambiguity is caused not by conflicting signal (such as that given by reticulation) but by sampling error. One clear example is the part of the

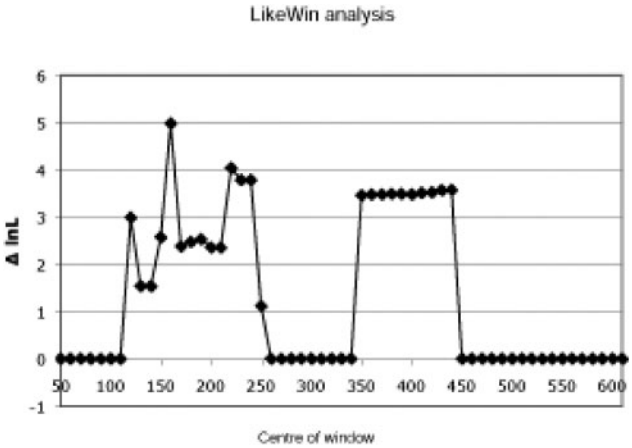

FIG. 9.—Difference in log likelihoods for sliding-window analysis of seven Salmonella *manB* sequences.

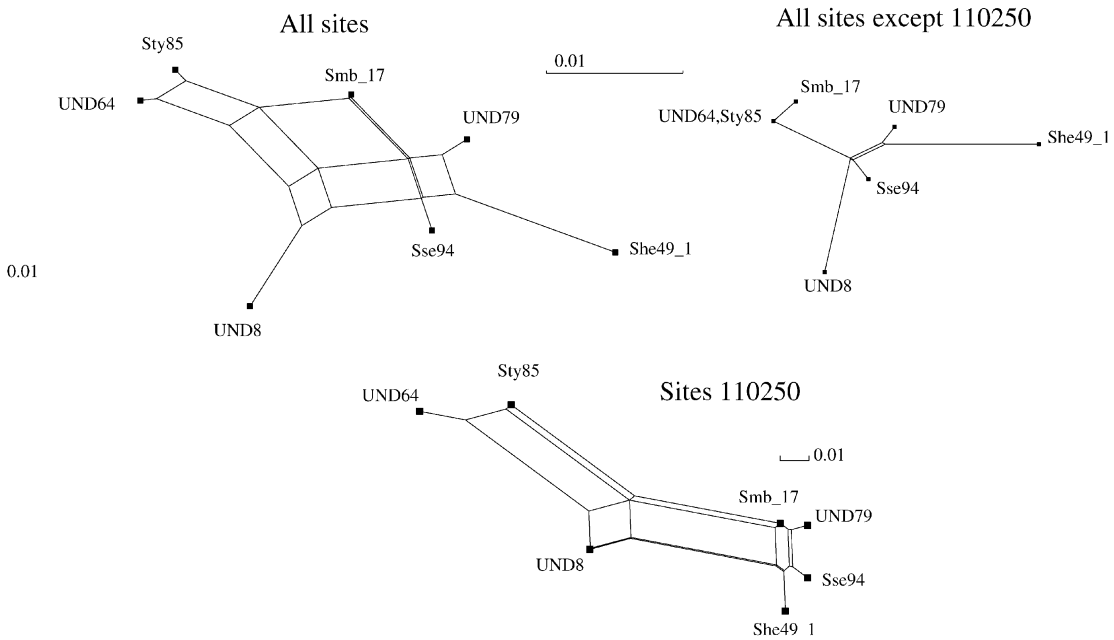

FIG. 10.—Networks produced for seven Salmonella *manB* sequences from different sites. Note the difference in scale in the lower network.

network near the connection point of the long branch leading to the Naron (76) taxon. These boxes represent ambiguity in the placement of the Naron taxa—an ambiguity already noted in phylogenetic analysis (Vigilant et al. 1991). If the taxon is deleted, so are the boxes. Thus Neighbor-Net is prone to long branch attraction, as are NJ and maximum likelihood. Unlike trees, however, networks can represent both the signal introduced by long branch attraction and the signal of the underlying phylogeny (Clements, Gray, and Choat 2002).

The Neighbor-Net did more than represent complexity and ambiguity. We also have a snapshot of the general structure in the data, a snapshot that is not restricted to a single bifurcating tree. The ''Africa 49'' cluster identified by Vigilant et al. (1991) (and validated by Penny et al. [1995]) is represented clearly. The divergence between African sequences is noticeably greater than the divergence between all other sequences. The African sequences are also more widely dispersed than the non-African sequences. Hence the Neighbor-Net is consistent with the analysis, and conclusions, of the original Vigilant et al. article, and this conclusion is reached without restricting attention to a single tree.

Archeal Chaperonin Data

Neighbor-Net analysis of the Archeal chaperonin sequences (fig. 12) rapidly detected the presence of conflicting signal. The Neighbor-Net is attempting to represent groupings resulting from gene conversion versus the separation between the two paralogs. The division between the α and β paralogs is clear, except for the position of *P. occultum* β. In the complete sequence, the signal grouping the two *P. occultum* taxa is stronger than that separating the α and β duplicates.

Although the method has detected conflict, it has not reconstructed the complete history. Neighbor-Net misrepresents some of the reticulation because it only constructs planar networks. It is not possible to group the α and β pairs for all three of *A. pernix*, *Pd. occultum* and *Pb. aerophilium* and still have a planar collection of splits. Indeed, grouping two of these pairs and splitting the paralogs would also violate planarity.

We therefore suggest that any reticulations detected by Neighbor-Net be investigated and validated with other, perhaps more detailed, methods. The advantage of Neighbor-Net is that it is rapid and scales well, producing a detailed overview of the entire data set. Other methods, like split decomposition and median networks, are not suited for analysis of larger data sets, but are still useful for detecting patterns in subsets of taxa.

In this case, we applied split decomposition to two subsets of the taxa (fig. 12). The split decomposition graph for *A. pernix, Pd. occultum* and *Pb. aerophilium* is non-planar, and represents both the grouping of the pairs of paralogs and the separation of the α and β sequences. Split decomposition does not detect any additional signal when applied to the other six sequences. The split decomposition graph constructed on the entire data set (not shown) is substantially less resolved than the Neighbor-Net.

The message is, then, to apply Neighbor-Net to the entire data set to rapidly obtain an overview of the structure and possible points of interest. Finer details can then be explored using more computationally intensive or detailed methods.

Discussion

Neighbor-Net rapidly provides a detailed snapshot of the data. The algorithm is an extension of the NJ method,

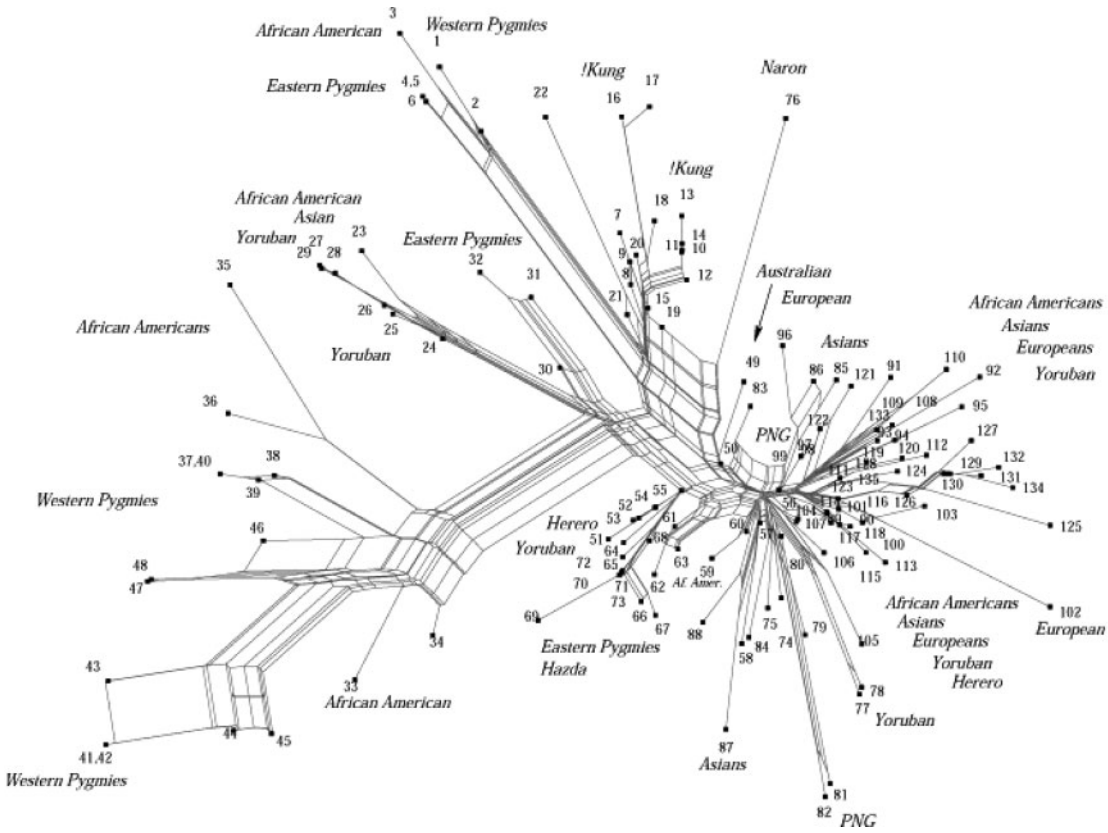

FIG. 11.—Neighbor-Net for the 135 human mitochondrial sequences studied by Vigilant et al. The numbering is identical to that of Vigilant et al.: Western Pygmies (1, 2, 37–48); Eastern Pygmies (4–6, 30–32, 65–73); !Kung (7–22); African Americans (3, 27, 33, 35, 36, 59, 63, 100); Yorubans (24–26, 29, 51, 57, 60, 63, 77, 78, 103, 106, 107); Australian (49); Herero (34, 52–56, 105, 127); Asians (23, 28, 58, 74, 75, 84–88, 90–93, 95, 98, 112, 113, 121–124, 128); Papua New Guineans (50, 79–82, 97, 108–110, 125, 129–135); Hazda (61, 62, 64, 83); Naron (76); Europeans (89, 94, 96, 99, 101, 102, 104, 111, 114–120).

using similar selection and reduction formulae. Unlike NJ, Neighbor-Net can represent conflicting signals in the data, whether they arise from sampling error or genuine recombinations.

Neighbor-Net is fast. The basic algorithm takes $O(n^3)$ time on n sequences, the same order of complexity as NJ. The major computational difficulty comes with edge weighting, which uses a least squares estimation under a non-negativity constraint. At present we use a combination of iterative techniques and combinatorial algorithms, but there is potential for substantial improvements in efficiency for this step. Even so, we have been able to relatively quickly (a few minutes) analyze data sets containing over 300 taxa on a 600 MHz laptop.

Neighbor-Net is consistent and, apparently, relatively efficient. We have proven consistency over a large class of distance matrices. However, as we saw with the archael chaperonins, the planarity constraint is not sufficiently general in some situations. In these cases, Neighbor-Net is not consistent, but neither are any tree-based methods. However, unlike tree-based methods, the network generally gives a clear indication of which parts of the network the complexity stems from, allowing us to focus in on those regions, with more detailed and computationally demanding methods.

Neighbor-Net is informative. Our three examples indicate that networks produced by Neighbor-Net are useful both as a representation of the overall structure of the data and as a guide for further analysis. A splits graph is a powerful representation tool, even if it does require some practice to interpret.

There remain many open questions. The most fundamental is the interpretation and validation of the splits graphs produced by Neighbor-Net. At present, we advocate use of the method as a technique for data representation and exploration, much in the same way as a scatter diagram can be used to explore the relationship between two real valued variables. To go beyond exploration to diagnosis we require a consistent framework for interpretation of splits graphs, particularly if we are to design meaningful significance tests. Recent progress toward solving these problems has been made by Bryant et al. (2003), who show that the splits in the network are estimations of the splits in the input trees. However, the interpretation is a little idealistic because it ignores the planarity constraint inherent in Neighbor-Net.

This leads us to the second shortcoming of the method—one highlighted by the Archael Chaperonin analysis. Neighbor-Net produces circular collections of splits. We concede that the definition of circular splits and circular distances is not biologically motivated. However, the key observation is that this ''mathematically motivated'' class of distance matrices includes treelike distances and the matrices generated by a large range of evolutionary

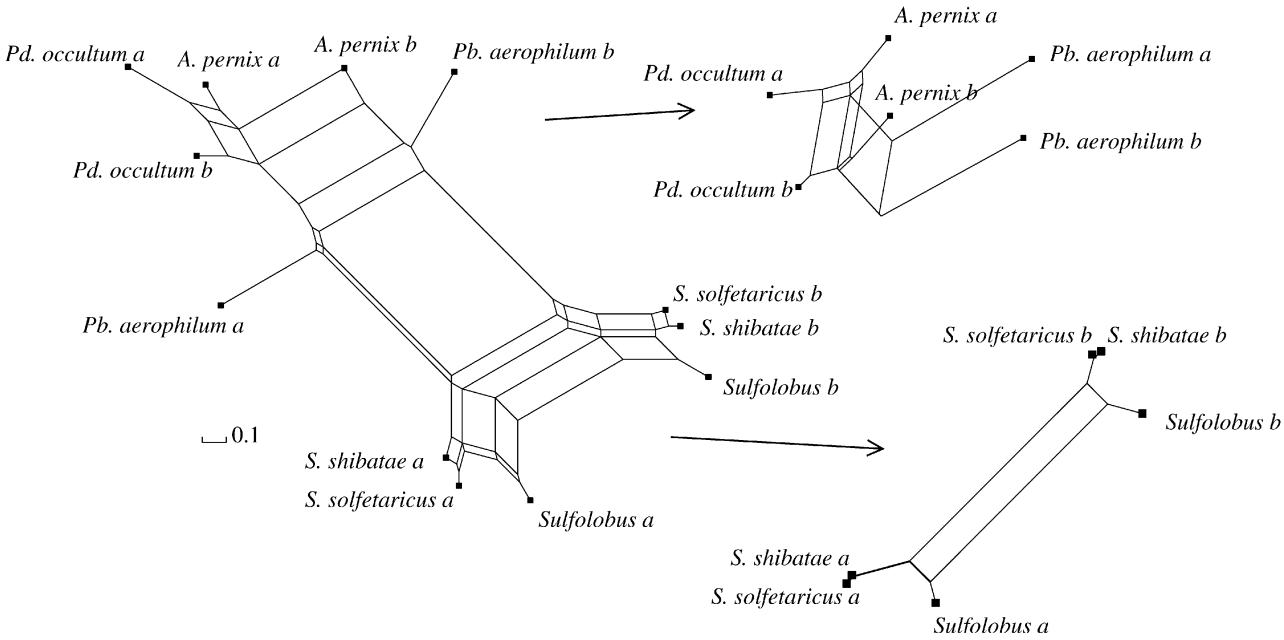

FIG. 12.—Neighbor-Net for the Archael chaperonins (left). Split decomposition graphs (right) for two subsets of the taxa set.

histories. The planarity property guarantees that the representation will not become so overly complicated as to be uninformative. Nevertheless, planar split graphs might not be general enough for some evolutionary histories. In these cases, Neighbor-Net can still give an indication of where such complexities arise, allowing one to focus on the relevant portions of the tree or network for more specialized analysis. There is still potential for network methods producing different but representative collections of splits.

Finally, we note that a splits graph is only one step toward a complete reconstruction of recombination histories. Under a standard evolutionary model, each gene or pair of contiguous segments has a treelike evolutionary history, and the network yields a composite of these different histories. The difficult problem of unravelling this composite history remains, although we have seen that Neighbor-Net provides a valuable first step.

Acknowledgments

We thank A. Sulakvelidze, D. Penny, and J. Archibald for providing the sequence data. This work was supported by NSERC grant 238975-01, FQRNT grant 2003-NC-81840 (D.B.), and The Swedish Research Council (V.M.).

Literature Cited

Archibald, J., and A. Roger. 2002*a*. Gene conversion and the evolution of euryarchaeal chaperonins: a maximum likelihood–based method for detecting conflicting phylogenetic signals. J. Mol. Evol. **55**:232–245.

Archibald, J., and A. Roger. 2002*b*. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. J. Mol. Biol. **316**:1041–1050.

Bandelt, H.-J., and A. Dress. 1992. A canonical decomposition theory for metrics on a finite set. Adv. Math. **92**:47–105.

Bandelt, H., P. Forster, B. Sykes, and M. Richards. 1995. Mitochondrial portraits of human population using median networks. Genetics **141**:743–753.

Bandelt, H., P. Forster, and A. Röhl. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. **16**:37–48.

Bruno, W., N. Socci, and A. Halpern. 2000. Weighted Neighbor Joining: a likelihood-based approach to distance based phylogeny. Mol. Biol. Evol. **17**:189–197.

Bryant, D., D. Huson, T. Kloepper, and K. Nieselt-Struwe. 2003. Distance corrections on recombinant sequences. Pp. 271–286 *in* G. Benson and R. Page, eds. WABI 2003: Algorithms in Bioinformatics, Third International Workshop, Proceedings. Lecture Notes in Computer Science 2812.

Bryant, D., and V. Moulton. 2003. Consistency of the NeighborNet algorithm for constructing phylogenetic networks. Technical report, School of Computer Science, McGill University.

Cavalli-Sforza, L., and A. Edwards. 1967. Phylogenetic analysis models and estimation procedures. Evolution **32**:550–570.

Clements, K., R. Gray, and J. Howard Choat. 2002. Rapid evolutionary divergences in reef fishes of the family Acanthuridae (Perciformes: Teleostei). Mol. Phylogenet. Evol. **26**:190–201.

Diday, E. 1986. Une représentation visuelle des classes empiétantes: les pyramides. RAIRO Automat.-Prod. Inform. Ind. **20**:475–526.

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

Fitch, W. 1997. Networks and viral evolution. J. Mol. Evol. **44**:S65–S75.

Gascuel, O. 1997*a*. Bionj: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

Gascuel, O. 1997*b*. Concerning the NJ algorithm and its unweighted version, UNJ. Pp. 149–170 *In* B. Mirkin, F.

McMorris, F. Roberts, and A. Rhetsky, eds., Mathematical Hierarchies and Biology, pages 149–170. AMS, Providence.
Grassly, N., and E. Holmes. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. **14**:239–247.
Huber, K., V. Moulton, P. Lockhart, and A. Dress. 2001. Pruned median networks: a technique for reducing the complexity of median networks. Mol. Phylogenet. Evol. **19**:302–310.
Huber, K., M. Langton, D. Penny, B. Moulton, and M. Hendy. 2002. Spectronet: a package for computing spectra and median networks. Appl. Bioinformatics **1**:159–161.
Huson, D. 1998. Splitstree—a program for analyzing and visualizing evolutionary data. Bioinformatics **14**:68–73.
Kotetishvili, M., O. Stine, A. Kreger, J. 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.
Ota, S., and W. Li. 2000. NJML: A hybrid algorithm for the Neighbor-Joining and maximum likelihood methods. Mol. Biol. Evol. **17**:1401–1409.
Penny, D., M. Steel, P. Waddell, and M. Hendy. 1995. Improved analysis of human mtDNA sequences support a recent African origin for Homo sapiens. Mol. Biol. Evol. **12**:863–882.
Posada, D., and K. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **14**:817–818.
Posada, D., and K. Crandall. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol. **16**: 37–45.
Saitou, N., and M. Nei. 1987. The Neighbor-Joining method: a new method for reconstruction of phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
Sawyer, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.
Strimmer, K., C. Wiuf, and V. Moulton. 2001. Recombination analysis using directed graphical models. Mol. Biol. Evol. **18**:97–99.
Swofford, D. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.
Templeton, A., K. Crandall, and C. Sing. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and dna sequence data. III. Cladogram estimation. Genetics **132**:619–633.
Vigilant, L., Stoneking, H. M. Harpending, K. Hawkes, and A. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. Science **253**:1503–1507.
Xu, S. 2000. Phylogenetic analysis under reticulate evolution. Mol. Biol. Evol. **17**:897–907.

William Martin, Associate Editor

Accepted September 5, 2003