Mitochondrial Portraits of Human Populations Using Median Networks

Hans-J. Bandelt,* Peter Forster,† Bryan C. Sykes‡ and Martin B. Richards‡

**Mathematisches Seminar der Universität Hamburg, D-20146 Hamburg, Germany, †Heinrich-Pette-Institut für Experimentelle Virologie und Immunologie an der Universität Hamburg, D-20251 Hamburg, Germany and ‡Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, United Kingdom*

Manuscript received February 25, 1995
Accepted for publication June 24, 1995

ABSTRACT

Analysis of variation in the hypervariable region of mitochondrial DNA (mtDNA) has emerged as an important tool for studying human evolution and migration. However, attempts to reconstruct optimal intraspecific mtDNA phylogenies frequently fail because parallel mutation events partly obscure the true evolutionary pathways. This makes it inadvisable to present a single phylogenetic tree at the expense of neglecting equally acceptable ones. As an alternative, we propose a novel network approach for portraying mtDNA relationships. For small sample sizes ($< \sim 50$), an unmodified median network contains all most parsimonious trees, displays graphically the full information content of the sequence data, and can easily be generated by hand. For larger sample sizes, we reduce the complexity of the network by identifying parallelisms. This reduction procedure is guided by a compatibility argument and an additional source of phylogenetic information: the frequencies of the mitochondrial haplotypes. As a spin-off, our approach can also assist in identifying sequencing errors, which manifest themselves in implausible network substructures. We illustrate the advantages of our approach with several examples from existing data sets.

ALTHOUGH the analysis of mitochondrial DNA (mtDNA) has emerged as an important tool in the study of human population structure and history (STONEKING 1993), the interpretation of mtDNA data is notoriously difficult. As for almost any molecular data set used for phylogenetic studies, human mtDNA data, whether comprising restriction fragment length polymorphism (RFLP) data or sequences from the hypervariable control region, almost invariably fail to form nested sets (or cliques *sensu* MEACHAM and ESTABROOK 1985) of haplotypes but exhibit incompatibility between pairs of characters. The reason for this is homoplasy (parallel mutation events or reversals); in conjunction with the low number of informative characters analyzed, this means that resolution is lost as the coalescent time is approached (HEDGES *et al.* 1991; MADDISON 1991).

Traditional tree-building methods, such as maximum parsimony (MP), maximum likelihood (ML), and distance methods, are therefore unsatisfactory when applied to human mtDNA data. For example, EXCOFFIER and SMOUSE (1994) calculated that the number of equally parsimonious trees for an RFLP data set of just 56 haplotypes exceeded one billion. In the light of this, the practice of presenting a single randomly chosen MP tree in which character conflicts have been arbitrarily resolved is likely to be as misleading as forming a strict consensus of two or more trees in which true polytomies may no longer be distinguished from ambiguities.

Problems may also arise with the application of distance or maximum likelihood methods. For instance, a method such as neighbor-joining (NJ) (SAITOU and NEI 1987) would produce a dichotomous tree unless estimated branch lengths become zero or negative. However, a typical mtDNA haplotype tree is notoriously polytomous at many branching points. Suppose, for example, that in a real tree the ancestral node of a clade that is supported by several mutations has many descendent branches, one of which carries three haplotypes that have each acquired a mutation in parallel with some distant haplotype branching off outside the clade. Such parallel events are easy to detect (and would be recognized as such by a parsimony method), but they affect the distance calculations so that the polytomy is artificially resolved by an earlier branching of that particular branch. The resulting spurious "ghost" link, though being of small length, would even be well supported by bootstrapping, provided no further parallelisms were involved. This is because the minor signal inducing the link is present whenever at least one of the three homoplasious sites is resampled. Most NJ trees (*e.g.*, in HORAI *et al.* 1993; MOUNTAIN *et al.* 1995) and also ML trees (see RESULTS) thus display an artificially high level of resolution when taken fresh out of the computer. Such ghost links could readily be discovered and contracted when one determines the most parsimonious assignments of mutations along the links of the tree. Postprocessing of computer outputs is however rarely done in the field (though see NERURKAR *et al.* 1993 for an exception), possibly because SWOFFORD and

Corresponding author: Hans-J. Bandelt, Mathematisches Seminar der Universität Hamburg, Bundesstrasse 55, D-20146 Hamburg, Germany.

OLSEN's (1990) remark still holds true: "Unfortunately, phylogenetic analysis is frequently treated as a black box into which data are fed and out of which 'The Tree' springs."

Progress has been made by discarding certain traditional views, for example, that there should be a single optimal solution in the form of a unique tree. EXCOFFIER and SMOUSE (1994) argue that "We need to begin thinking about classes of acceptable solutions that allow us to bracket our estimates and our ignorance." They employ simple (one- or two-step) networks as an intermediate stage of their analyses, but they nevertheless still prefer to present their conclusions in the form of a set of (near-) optimal trees. An early attempt to visualize parallelisms in the form of reticulations was proposed in unpublished work and lectures by the zoologist UDO REMPE (University of Kiel, Germany) in the early 1980s. Occasional usages of reticulate networks in the literature are listed in BANDELT (1994). Another important development is the incorporation of frequency in the reconstruction process for intraspecific phylogenies (see EXCOFFIER and LANGANEY 1989; CRANDALL and TEMPLETON 1993; TEMPLETON 1993).

We argue that mtDNA data are best analyzed by a strict network approach that distinguishes between unresolvable and resolvable character conflicts, leaving a compact intelligible representation of plausible solutions. Our approach is to work with median networks (GUÉNOCHE 1986; BARTHÉLEMY 1989; BANDELT 1992) generated by partitioning the groups of haplotypes character by character, as we describe below. We show that unmodified median networks are guaranteed to include all most parsimonious trees. Unlike tree-building methods or frequency distributions of pairwise comparisons, our network approach can highlight character conflicts in the form of reticulations, which can then be interpreted in terms of homoplasy (high rates of homoplasy might even lead to a single haplotype being independently derived along different routes from the same ancestor), recombination, sequence error, or superimposed sequences. We indicate when a reticulation in the network may justifiably be resolved on a compatibility basis in conjunction with a frequency-based argument and thus exhibit likely evolutionary routes through a network. We present networks for three human populations with dissimilar mitochondrial phylogenetic structures to show that networks provide a much more useful and informative mitochondrial portrait of the populations concerned than can be obtained from more traditional approaches.

MATERIALS AND METHODS

Preprocessing the data: Consider the case that mtDNA of N individuals from a population is examined. The mtDNA haplotypes would typically be either patterns of restriction sites or ~300-bp sequences of DNA within the first hypervariable segment of the control region. Restriction haplotypes constitute genuine binary data since a site is either present or absent, but a site of a DNA sequence can be any one of the nucleotides A, G, C, T. Fortunately, the vast majority of observed changes are transitions (A ↔ G or C ↔ T), so that mtDNA sequences may be turned into binary data in the following Procrustean fashion. First, one would stipulate that an observed transversion at a site was a single event since transversions are by at least an order of magnitude less likely than transitions (VIGILANT 1990; WARD *et al.* 1991). In the rare case that all four nucleotides are found at a certain site of the sampled DNA sequences, we would then be left with four possibilities for the transversion event (A ↔ C, A ↔ T, G ↔ C, G ↔T), among which usually one is favored by invoking the parsimony criterion. In this manner, every site at which more than two nucleotides are observed in the sampled sequences is regarded as a pair or a triplet of binary characters, one of which denotes a transversional change. The thus modified haplotypes can now be handled in the same way as restriction haplotypes. Second, in case few nucleotides are scored ambiguous or deleted, we would omit those sequences or sites for the primary analysis and then fit the omitted sequences or sites to the obtained network so that the number of additional nodes and links is minimized.

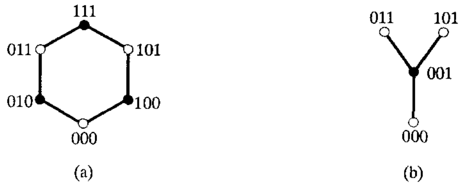

FIGURE 1.—Two connected (one-step) subnetworks of the three-dimensional cube that include the three vectors 000, 011, 101.

The next step is to eliminate multiple rows or columns from the processed data matrix. Identical haplotypes are pooled, leaving n distinct haplotypes. Then the number of individuals with haplotype i is the frequency f_i. Sites that are constant among the haplotypes are removed, so that henceforth the sites are all variable and each site of the n distinct haplotypes splits the haplotypes into two groups according to the two states at the site. Each split is coded as a binary character with states 0 and 1 (under the convention that a certain reference haplotype receives all character states 0, without imposing polarity). We now pool a set of sites that give rise to the same split into one character. We then obtain k distinct characters, where the weight w_i of character i is the number of sites that give rise to this character. The haplotypes are thus represented by 0–1 vectors of length k. The information on character weights and haplotype frequencies is irrelevant for constructing the intermediate nodes of the median network but is used later on in the reduction procedure and the final graphical display.

Definition of the median network: After preprocessing we have n 0–1 vectors of length k. These vectors can be thought of as being located in the k-dimensional cube of all 2^k possible 0–1 vectors of length k. This hypercube is the "one-step network" of the $2k$ distinct 0–1 vectors in the sense that two vectors are linked exactly when they differ in a single coordinate. In contrast, the one-step network of only the given n vectors need not be connected; consider, for instance, the three vectors 000, 011, 101. Adding an intermediate vector for each pair to connect the three vectors yields a six-cycle, which can be realized within the three-dimensional cube by 000, 010, 011, 111, 101, 100 (Figure 1a). A more parsimonious connection of the three vectors is obtained by adjoining a single vector, viz., 001 (Figure 1b). This vector is the "consensus" vector (or "median" vector *sensu* FARRIS 1970) of 000, 011, and 101; it is computed componentwise by applying the majority rule. This example suggests that intermediate vectors can be reconstructed as follows. Departing from the initial set of n vectors, we successively enlarge the set of already constructed vectors by attaching the consensus vector of any triplet of vectors constructed so far. The process terminates with a set of vectors that includes all consensus vectors of its triplets. The one-step network of this set is necessarily connected (since all k characters are different) and will be referred to as the median network generated by the n given vectors. This network is independent of the order in which the consensus vectors were generated.

Most parsimonious trees in median networks: In the following we demonstrate that the median network generated from given 0–1 vectors harbors all most parsimonious trees. In the ideal case that no parallelisms or reversals occurred in the evolution of the n haplotypes, the median network generated by the corresponding vectors reconstructs precisely the undirected tree along which the haplotypes evolved, which is then obviously the unique most parsimonious tree. In the general case, for any three 0–1 vectors the consensus vector constitutes the branching node of the unique most parsimonious tree that connects them. Uniqueness follows from the evident fact that for three vectors a most parsimonious tree is free of homoplasy. A median network thus includes the consensus vector and hence the most parsimonious tree for each triplet of its nodes. This proposition can be extended to more than three nodes, viz., a median network of 0–1 vectors is known to include at least one consensus vector for every subset of vectors (which may be weighted according to frequency) (cf. BANDELT and BARTHÉLEMY 1984). If the consensus vector is not unique (as may only happen when the total number of individuals is even), then a consensus vector belonging to the median network is obtained by resolving ties in favor of an arbitrary haplotype x. This vector is then the consensus vector in the median network that is closest to the distinguished haplotype x and can be regarded as the unique consensus vector with respect to the modified haplotype distribution where the frequency of x is increased by 1.

A median network includes at least one most parsimonious tree for every set of its vectors according to VAN DE VEL (1993). This fact can be regarded as a direct consequence of the fundamental "retract" property of median networks: given a median network M generated by some vectors within a hypercube H all links of which are of unit length, there exists a mapping φ from the node-set of the hypercube to the network M such that every node of M is left fixed, and two linked nodes of H are sent to either two linked nodes or the same node of M (BANDELT 1984). As a consequence, the retraction mapping φ sends any path with p links to a path in M with at most p links. In particular, the distance in H between two nodes of M can be read off from M as the length of a shortest path within M. Thus, for any set V of nodes from M, every subnetwork T of H constituting a most parsimonious tree for V is sent by φ to a most parsimonious tree $\varphi(T)$ that lies entirely in the median network M. Indeed, $\varphi(T)$ is connected and has the same number of links as T. Consequently, $\varphi(T)$ does not include any loops or cycles, whence no two distinct nodes of T are mapped to the same node of $\varphi(T)$. We have therefore proven the following result in the unweighted case: all most parsimonious trees for a set V of 0–1 vectors can be realized in the median network M generated by V. Here a most parsimonious tree T is understood to have no superfluous links, so that T can be realized within the hypercube H. The case of weighted coordinates with all weights being positive integers can quite easily be reduced to this unweighted situation by splitting each coordinate i of weight w_i into w_i sites that only change in concert. For a different proof, see VAN DE VEL (1993).

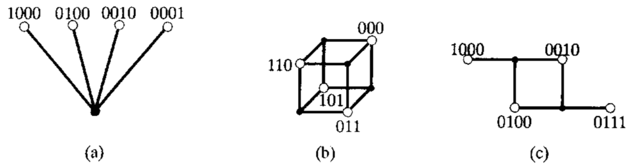

FIGURE 2.—Three different sets of four 0–1 vectors and the median networks they generate.

The network provides a much more informative picture of the data than a mere strict consensus of all most parsimonious trees. To give an illustration, consider the following three sets of 0–1 vectors: (a) 1000, 0100, 0010, 0001; (b) 110, 101, 011, 000; (c) 1000, 0111, 0100, 0010. The median networks they generate are shown in Figure 2. In each case the strict consensus of all most parsimonious trees is a polytomous tree, yet the data sets differ considerably in structure. The biggest contrast is between (a) and (b); although all pairwise distances equal 2, we have a true polytomous tree in the first case, whereas for (b) we have a totally ambiguous situation, implying a considerable amount of homoplasy. This distinction is also emphasized by BANDELT and DRESS (1992). In the third case there is still some ambiguity in the choice of a best tree but to a lesser extent than in (b); in particular, a tree with an inner link separating 1000 and 0111 from 0100 and 0010 would not receive support from the data.

Construction of the median network: The definition alone does not lend itself to a practical procedure for constructing the median network from a set of given vectors since iteratively generating consensus vectors from triplets of already constructed vectors would be too tedious. Instead, one can easily produce the network by processing the characters one after another: in the jth step one would then obtain the median network generated by the truncated vectors consisting only of the first j components. The final network is independent of the actual order in which the characters are processed. The initialization for $j = 0$ sets M_0 to be the trivial network with a single node representing all n haplotypes. The recursive step of the construction is as follows. Assume we have already represented the first $j - 1$ components of the given vectors by their median network M_{j-1} within the $(j - 1)$-dimensional cube. The jth character (being the jth component of the vectors) splits the haplotypes into two complementary groups A and B, where A comprises the haplotypes with jth component 0 and B those with 1. Next we associate each of the remaining nodes of M_{j-1} to either one of the groups A and B, or to both groups, depending on whether a node can be regarded as an intermediate vector for group A or B. The sets A and B are thus extended to larger sets $[A]$ and $[B]$ that together cover all nodes of M_{j-1}. A vector x from M_{j-1} belongs to $[A]$ exactly when for each index i with $1 \le i < j$ its ith component x_i matches the corresponding component of at least one vector of A; formally,

$$[A] = \{x \mid x \text{ is a node from } M_{j-1} \text{ such that } x_i \in A_i \text{ for all } 1 \le i < j\} \quad (1)$$

where A_i comprises the ith components of the vectors from A (and thus equals either $\{0\}$, $\{1\}$, or $\{0,1\}$). In the same way $[B]$ is defined. By our initial preprocessing all k characters induce different splits, and, consequently, $[A]$ and $[B]$ must intersect in at least one node. This common intersection constitutes a median subnetwork in its own right as it equals the intersection of M_{j-1} with a hypercube (formed by all 0–1 vectors x of length $j - 1$ with $x_i \in A_i \cap B_i$ for $1 \le i < j$). Now,

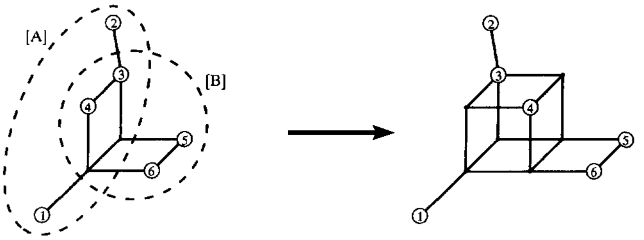

FIGURE 3.—Expansion of a median network (generated by nodes 1, 2, 3, 4, 5, 6) by introducing the new split $A = \{1, 2, 3\}$, $B = \{4, 5, 6\}$.

introducing the jth components, we will attach to each vector from $[A]$ the jth component 0 and to each vector from $[B]$ the jth component 1, so that for each vector from $[A] \cap [B]$ we obtain two new vectors that differ only in the jth component and are therefore linked in the new network M_j. All these links are associated to the jth character. Figure 3 illustrates this expansion procedure. The network M_j obtained in this step is the median network generated by the first j components of the n given vectors. That M_j is indeed a median network can readily be checked by considering the consensus vectors of all triplets drawn from M_j. After k steps, *i.e.*, when all characters have been processed, one arrives at the desired median network (BANDELT 1992).

For the manual construction of the network one can dispense with labeling the nodes by 0–1 vectors since the splits induced by the characters processed so far can directly be read off from the network. The split associated to a link of a median network can be determined as follows. Label the two nodes x and y of this link by 0 and 1, respectively; then all nodes that are closer to x than to y receive label 0, while all the other nodes receive label 1, thus establishing the split. It suffices, of course, just to determine the links that connect the 0-labeled part with the 1-labeled part. To this end, scan all hypercubes in the network that include the nodes x and y, mark all links in each of these hypercubes that correspond to the same dimension as the link between x and y, and continue this marking process with the newly marked links until no links can be marked further. Then the marked links constitute the links that connect the two parts of the split associated to the link between x and y.

Median networks generated by even a few vectors can become very large if the amount of homoplasy is very high. For instance, five vectors could generate a network that would require as many as 76 intermediate nodes, and for six vectors one would need 2640 additional nodes in the worst case (cf. BANDELT and VAN DE VEL 1991). It would not be convenient to draw even the former network with its 81 nodes since its inherent symmetry could only be well appreciated in five-dimensional space. Fortunately, one is much better off with mtDNA sequence data in practice, although for more than, say, 30 haplotypes, the resulting median networks would become too large for display and could only be stored in the computer.

Network reduction: The main goal in presenting the mtDNA variation in a population sample is to have an intelligible representation of probable evolutionary pathways. To focus on the most probable pathways in the network, we can choose to resolve reticulations by identifying obvious parallelisms. To this end the weight of a character (representing the number of corresponding sites) and the frequencies of haplotypes constitute essential information for the decisions to be made. The criterion we use utilizes compatibility (MEACHAM and ESTABROOK 1985) and is guided by the fact that mutational events would typically proceed from the more frequent haplotypes to the less frequent ones (for a justification from coalescent theory see DONNELLY AND TAVARÉ 1986; EXCOFFIER *et al.* 1992; CASTELLOE and TEMPLETON 1994).

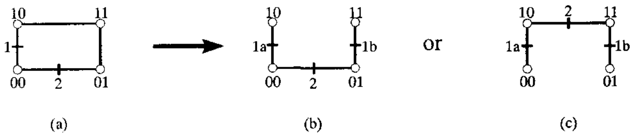

FIGURE 4.—The two evolutionary pathways supported by a pair of incompatible characters where the first character has (much) lower weight than the second one.

To begin with, consider two incompatible (binary) characters: then all four combinations 00, 01, 10, 11 of states are present among the haplotypes (Figure 4a). If the second character has a considerably larger weight than the first one, $w_2 \geq 2w_1$ say, then it is more likely that the first character has undergone an additional mutation. This would suggest two plausible pathways: either the unique mutation in character no. 2 occurs between 00 and 01 (Figure 4b), or this step is between 10 and 11 (Figure 4c). Frequency information can assist in choosing between the two alternatives. Suppose combinations 00 and 01 are more frequent among the individual sequences than 10 and 11, respectively, then the first solution (Figure 4b) is more likely than the second one. This is the kind of operation that we would perform to eliminate one rectangle (or a ladder of rectangles as in Figure 5 below) of a median network. Moreover, with regard to only the two characters under consideration, the criterion (based on haplotype frequencies) of EXCOFFIER and SMOUSE (1994) would select the same alternative. The strength of the argument increases with the weight ratio w_2/w_1 and the frequency surplus $f(00) + f(01) - f(10) - f(11)$. The choice between the two alternatives in Figure 4 would be made easier if character no. 2 gets support from further characters. Specifically, assume that we find (compatible) characters, numbered 2 to $k + 1$, with associated splits $\{A_i, B_i\}$ $(i = 2, \ldots, k + 1)$ satisfying

$$A_2 \subseteq A_3 \subseteq \cdots \subseteq A_{k+1}, \tag{2}$$

that is, these k characters describe an evolutionary path of haplotypes. Assume further that character no. 1 with split $\{A_1, B_1\}$ is incompatible with each of the other characters, so that

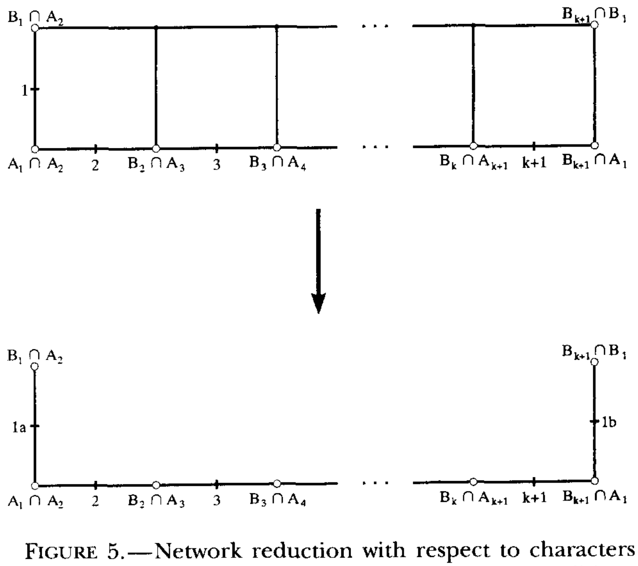

FIGURE 5.—Network reduction with respect to characters with splits A_i, B_i $(i = 1, \ldots, k + 1)$ meeting the conditions (2)–(7): character no. 1 is divided into two independent characters (1a and 1b).

$$A_1 \cap A_2,\quad B_1 \cap A_2,\quad B_{k+1} \cap A_1,\quad B_{k+1} \cap B_1 \neq \emptyset. \quad (3)$$

The median network generated by the $k + 1$ characters is thus a ladder comprising k rectangles. Now, suppose the extreme case that all inner nodes of one side of the ladder do not represent observed haplotypes, say

$$B_1 \cap B_2 \cap A_{k+1} = \emptyset \quad (4)$$

holds true, as is indicated in the ladder of Figure 5. We postulate two changes in character no. 1 if its weight is at most one-half of the total weight of the clique:

$$\sum_{i=2}^{k+1} w_i \geq 2w_1. \quad (5)$$

Recall that the weighting scheme handled here would simply define the weight of a character as the number of sites pooled in this character, except that to block the resolution of a conflict between a single transversional event and a pair of compatible transitional changes, we would give each transversion a weight of 1.5. We eventually adopt the evolutionary pathway indicated in Figure 5, connecting the haplotypes from $B_1 \cap A_2$ with those from $B_1 \cap B_{k+1}$ provided that the frequencies of the haplotypes from the $k + 1$ intermediate sets $A_1 \cap A_2$, $B_2 \cap B_3$, $B_3 \cap B_4, \ldots, B_k \cap A_{k+1}$, $A_1 \cap A_{k+1}$ are not too low compared to those in $B_1 \cap A_2 = B_1 \cap A_{k+1}$ and $B_1 \cap B_{k+1} = B_1 \cap B_2$. We specify this by requiring that strict majorities of the individuals allocated to A_{k+1} and B_2, respectively, belong to A_1, that is,

$$f(A_1 \cap A_{k+1}) = f(A_1 \cap A_2) + f(B_2 \cap A_{k+1}) > f(B_1 \cap A_2), \quad (6)$$

$$f(A_1 \cap B_2) = f(A_1 \cap B_{k+1}) + f(B_2 \cap A_{k+1}) > f(B_1 \cap B_{k+1}), \quad (7)$$

where for any set S of haplotypes, $f(S)$ denotes the number of individuals whose haplotypes are in S. Under these circumstances character no. 1 is thus substituted by two new characters, coded by 1a and 1b, with splits $\{B_1 \cap A_2, A_1 \cup B_2\}$ and $\{B_{k+1} \cap B_1, A_{k+1} \cup A_1\}$, respectively. The reduction process is now executed as follows. For each character j we determine the maximal sets of k characters that are incompatible with character j, such that conditions (2)–(7) are met (where character j plays the role of the first character and where k depends on j). The reduction step is then performed for the character set meeting these conditions and yielding the maximum weight ratio $(w_2 + \cdots + w_{k+1})/w_1$. Ties are broken in favor of larger frequency surplus $f(A_1 \cap A_{k+1}) + f(A_1 \cap B_2) - f(B_1 \cap A_2) - f(B_1 \cap B_{k+1})$ in inequalities (6) and (7). The data obtained thus (with one character substituted by two compatible new characters) are subjected to the same analysis. The process continues until no further character substitution is possible according to our rules. The median network generated by the resulting modified data is then referred to as the reduced network.

Graphical display: In addition to specifying mutational events, our median networks also display frequencies of haplotypes and weights of characters. Each node represents either an observed haplotype (drawn as a circle into which the code number of either the haplotype or the corresponding individuals are written) or a hypothetical intermediate haplotype (denoted by a dot or a small open circle). The area of the circle representing haplotype i is proportional to the number f_i of individuals with that haplotype. The lengths of the links corresponding to character i are proportional to its weight w_i (which is the number of sites that were pooled in character i if we do not weight sites); it suffices to label just one of the links by the sites associated to character i. We specify nontransitional mutations (transversions, insertions, deletions) at sites by letters or symbols following the site number. In the case of inferred multiple hits, the distinct mutational events at a single site are distinguished by suffixes a, b, c, etc. To recover the original data matrix from this graphical representation, one would have to further specify the actual sequence of one node. This is conveniently done by marking one node that serves as a consensus sequence and recording the sites at which it differs from a (published) reference sequence. In this fashion we obtain a compact iconic encoding of the data, requiring no additional tabular information.

Computation of most parsimonious trees: In contrast to the nonreduced median network, there is no formal guarantee that the reduced median network contains (all) MP trees, but in practice this seems to be the case. Unfortunately we cannot rely on PAUP (version 3.1.1., SWOFFORD 1993) for an appropriate count of MP trees (see RESULTS). First, the branch-and-bound option omits MP trees. Second, it counts redundant trees: two MP trees should be considered different only if no successive collapsing of links in one tree could ever retrieve the other tree. Briefly, the most parsimonious solutions should consist only of minimally resolved and thus maximally polytomous MP trees. Otherwise, we would retain redundant information: if tree T_2 resolves tree T_1 (in that all splits of T_1 are also found in T_2 but not vice versa), then any coalescence supported by tree T_2 would also be supported by tree T_1. Third, for $> \sim 20$ haplotypes, there is no guarantee to obtain any MP trees in a reasonable time span. Any exact parsimony procedure tailored to mtDNA population data would profit from applying partitioning strategies along the lines of HENDY *et al.* (1980). Even some manual preprocessing may reduce the number of haplotypes considerably before the application of an implemented exact algorithm. The most trivial case is the following: assume there is a character j with the property that all but one haplotype y possess the same state; then every MP tree assigns y to a terminal node. Storing this information, we may subsequently dismiss character j and substitute y by its neighboring node z (cf. FITCH 1977). If z also represents an observed haplotype, we would thus have reduced the effective number of haplotypes by one. Notice that z constitutes a cut node of the median network, that is, removing z would disconnect the network. More generally, consider a situation where a cut node z (*e.g.*, node * in Figure 6) occurs anywhere in a median network M. Then the haplotypes different from z can be partitioned into two sets X_1 and X_2 such that all paths in M from haplotypes of X_1 to those of X_2 pass through z. The network M is then the union of the median networks M_1 and M_2 generated by $X_1 \cup \{z\}$ and $X_2 \cup \{z\}$, respectively, such that M_1 and M_2 only intersect in the node z. In such a case, no character can correspond to both a link of M_1 and a link of M_2. Hence the characters are partitioned into the nonempty sets I_1 and I_2 such that all nodes of X_2 have the same state for each character from I_1 and all nodes of X_1 have the same state for each character from I_2; this is an equivalent way of expressing that the median network M has a cut node. Since all MP trees for the set X of observed haplotypes occur within the median network M, they are exactly the trees composed by the MP trees for $X_1 \cup \{z\}$ and $X_2 \cup \{z\}$, respectively. In a typical application of this decomposition principle to medium-sized data sets, $X_2 \cup \{z\}$, say, would be fairly small and generate a (star-like) tree, while the major component M_1 generated by $X_1 \cup \{z\}$ would be treated further until all pending subtrees are removed. Yet another rule allows reduction of the number of haplotypes in the remaining major component: partition the actual haplotypes into two sets, Y and Z, such that at least one of the MP trees for Y includes all haplotypes from Z as (inner) nodes (where the reduced median network for the actual haplotypes may assist in guessing a candidate Z). If such a set Z has been found, then the MP trees for $Y \cup Z$ are exactly the MP trees

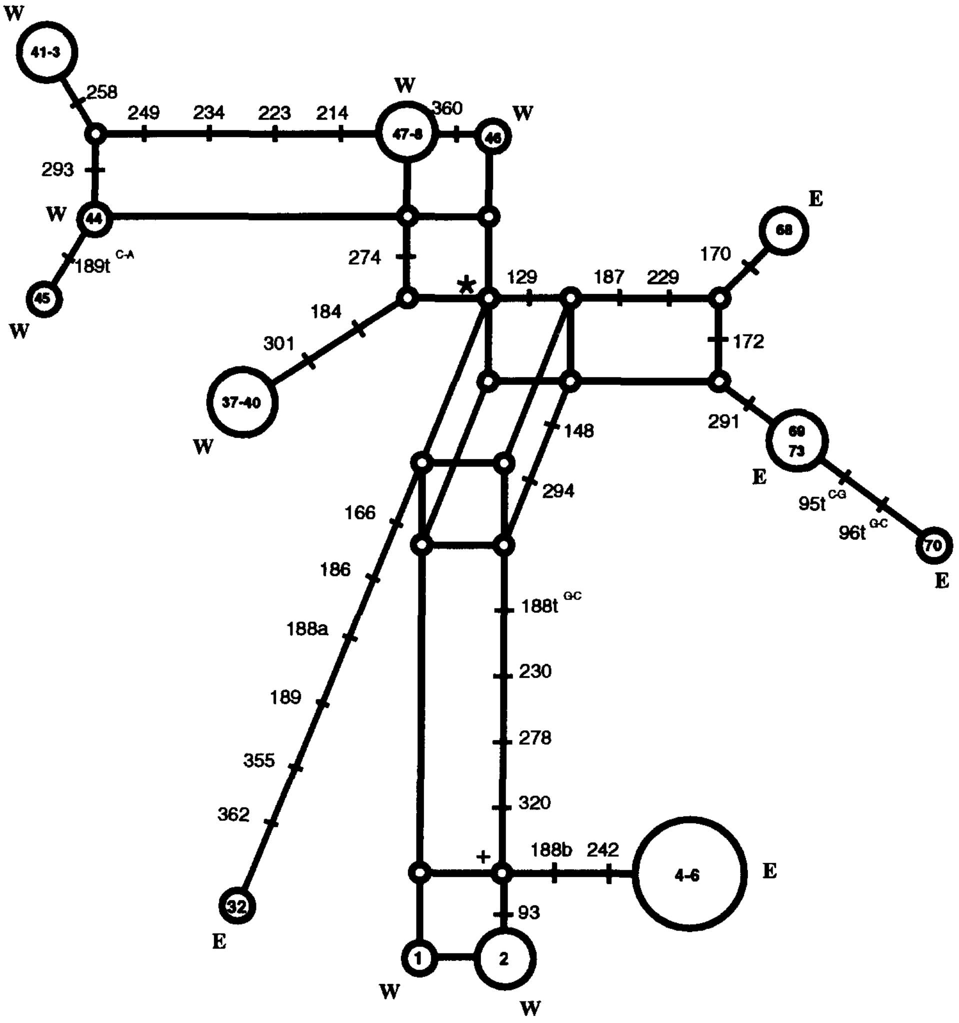

FIGURE 6.—The median network (containing all nine most parsimonious trees) for 35 West and East Pygmies, denoted W and E, respectively (data from VIGILANT 1990). The node marked by * represents one of the two consensus sequences, which deviates from the reference sequence at sites 129, 187, 189, 223, 278, 294, 311. Several non-Pygmy sequences (not shown) are located in the vicinity of the node marked by +.

for *Y* that include *Z* as well. The preceding two "pruning" operations, decomposition along cut nodes and deletion of candidate intermediate haplotypes, should suffice to cut down data sets of up to ~40 haplotypes into pieces that are amenable to exact parsimony algorithms.

Error detection: Two common types of systematic sequencing errors often leave their mark in networks in the form of reticulations. The first type is a reference bias error that arises when variant sites are overlooked, either when reading the sequencing output or during subsequent documentation procedures. If the overlooked variant is shared by other haplotypes in the data set and if this mutation is associated to an internal rather than a peripheral link in the network of the data set, then a reticulation will result, with the erroneous site defining one link of the reticulation. The second type of error is a baseshift error that arises when misjudging the position of a variant base on the sequencing output or misassigning a variant to a column of the data matrix. Baseshift errors create reticulations under the same conditions described above, but in this case the reticulation is more striking because a link with an adjacent (baseshifted) site emanates from the reticulation (cf. haplotype 17 in Figure 7). Nevertheless, as most reticulations in a mitochondrial network are due not to errors but to homoplasy, we recommend the following criteria to distinguish artefactual reticulations from genuine ones. Since the same error is unlikely to be committed more than once, an erroneous haplotype should be unique and without derivatives in the network. Furthermore, nodes with numerous derivatives are quite likely to give rise to genuine reticulations, since the more derivatives a node has the more likely it is that some of these derivatives are due to parallel mutation events. It is therefore generally not profitable to question these reticulations. However, reticulations toward the periphery of the network may be more suspect and merit rechecking.

RESULTS

The population portraits presented here concern data from the mtDNA control region between nt 16000 and nt 16400 of the reference sequence (ANDERSON *et al.* 1981). For convenience, we refer to variant bases at these sites by omitting the 16 prefix; hence, for example, a transition from the reference sequence at nt 16189 is marked 189. Transversions are represented by the suffix t. We draw on published sequence data for the following populations: Pygmies (Zaire/Central African Republic), Nuu-Chah-Nulth (Canada), and Frisians (Germany/Netherlands). Briefly, we propose candidate sequences for the human mitochondrial coalescent, we dismiss the rationale for discerning only four native American haplogroups, and we identify the main founding haplotype of European mtDNA.

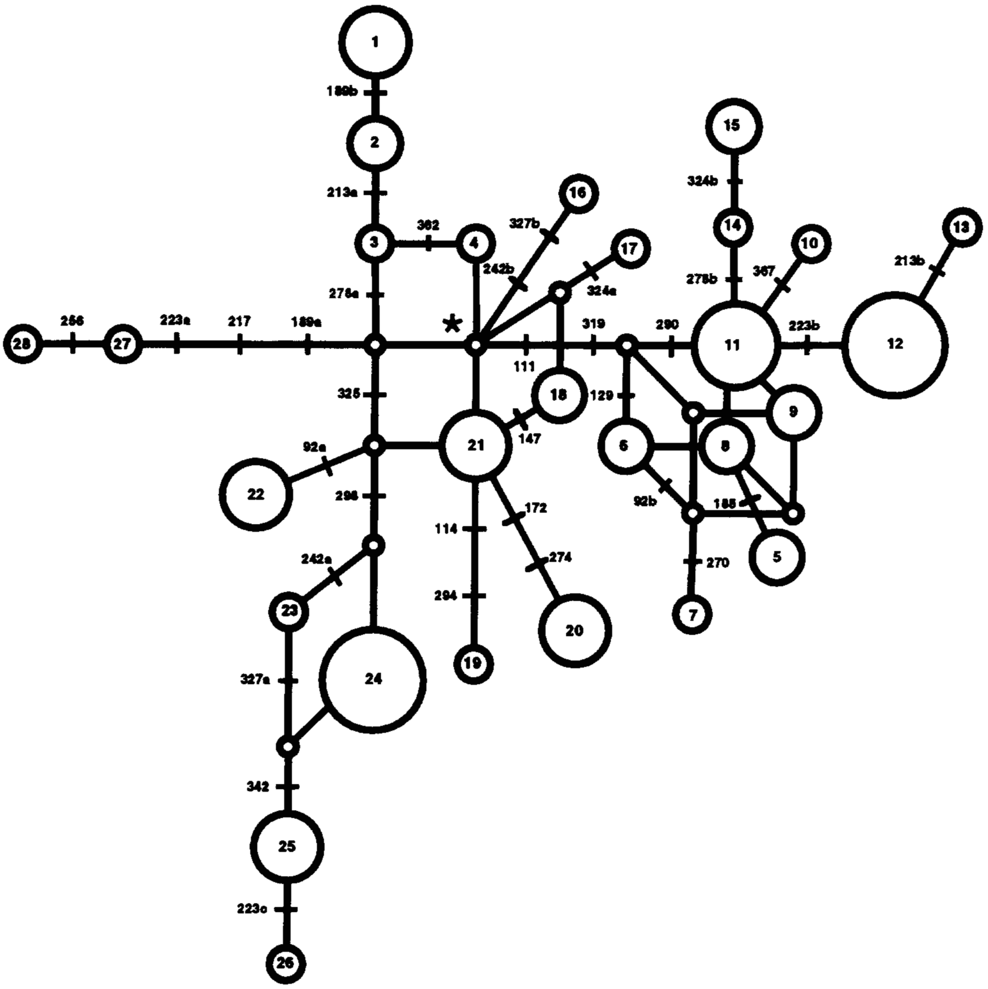

FIGURE 7.—The reduced network (containing 896 most parsimonious trees) for 28 different haplotypes among 63 Nuu-Chah-Nulth (data from WARD *et al.* 1991). The node marked by * represents the consensus sequence, which deviates from the reference sequence at sites 223, 362.

Figure 6 shows the median network for 35 complete control region sequences from two Pygmy populations analyzed for the region between 90 and 365 (data from VIGILANT 1990). The node marked by * represents one of the two consensus sequences. Sites 129, 172, 293 each conflict with a large number of sites making it very likely that they are homoplasious, but it is inadvisable to decide on the most likely tree in the network on the basis of the present data set, since the distribution of haplotypes is too sparse. The homoplasy is reflected by the coexistence of nine most parsimonious trees, which incidentally together cover all nodes of the network. The MP trees are readily determined by inspection of the network: they are composed of three MP trees for sequences 37–48 plus the consensus sequence *, and three MP trees for the remaining 10 sequences plus * (see MATERIALS AND METHODS for the decomposition argument). The result would be the same if we disregarded the four-state site 188: on each of the nine MP trees (for the other sites) site 188 would just need the necessary three changes and no more. Therefore the preprocessing was indeed most parsimonious. Surprisingly, the branch-and-bound option of PAUP omits three MP trees, and more confusingly, two of the remaining six trees are presented in the form of two triplets of topologically identical trees.

Empty nodes in the network often predict the existence of haplotypes. We can verify some of these predictions by fitting in the seven incomplete Pygmy sequences of the VIGILANT data: sequences 30, 31 could be allocated to two empty nodes in the link leading to sequence 32, whereas sequences 65, 71, 72 would occupy available nodes in the cluster formed by sequences 68, 69, 73 in the network. Only the incorporation of sequence 66 and its apparent neighbor sequence 67 would necessitate further reticulation, which would however be amenable to our reduction procedure, leaving only one additional square. The consensus node * must be very ancient (about five to seven transitions old, corresponding to at least 100,000 years according to the transition rate estimated by VIGILANT 1990). Interestingly, typical haplotypes of geographically diverse populations described by VIGILANT (1990) and DI RIENZO and WILSON (1991) (Yoruban, !Kung, Herero, Middle Eastern, Eurasian, Papuan, and Australian) seem to branch out from the network cube defined by the consensus node * and the seven sequences deviating from * at sites 129, 172, and 294. This suggests that the sequence of the human mitochondrial coalescent is in the immediate vicinity of this cube. The network topology clearly reflects the genetic difference between East and West Pygmies. An apparent exception could be perceived in the existence of the unspecific group containing both West and East Pygmies that cluster closely (within three transitions) around the node marked by + in Figure 6, along with members of diverse populations: two African Americans (GREENBERG *et al.* 1983, sample 3 in VIGILANT 1990), one Ugandan (100 in HORAI and HAYASAKA 1990), and intriguingly, one Sardinian and one Middle Easterner (22 and 55 from DI RIENZO and WILSON 1991). Whereas the Sardinian and the Middle Easterner represent clear outliers in their respective geographic regions, they fit perfectly in the cluster descended from node +. The apparent paradox of several ethnically specific Pygmy branches and one very unspecific cluster might be explained by postulating that the unspecific cluster in fact represents a new African population that diverged from the East Pygmy population two to six transitions ago and that has dispersed, at low frequency, into East and West Pygmies, into the Mediterranean, and into the ancestors of the African-American population. This new population may conceivably constitute a part of the Bantu population, whose present dispersal due to immigration into southern Africa and slavery is historically recorded. When sufficient Bantu mtDNA sequences become available, it will be interesting to test this interpretation.

Figure 7 shows the reduced network for 28 haplotypes from 63 individuals from the Nuu-Chah-Nulth of Vancouver Island (data of WARD *et al.* 1991) between sites 24 and 383. Although in the following we use this specific publication to demonstrate weaknesses of traditional phylogenetic analyses, our observations would apply to many examples in the literature. (1) The most striking feature of the reduced network is the finding of a fifth major haplogroup among native Americans if one superimposes the haplogroup definitions of HORAI *et al.* (1993) and WARD *et al.* (1991) (cf. BAILLIET 1994); group II of HORAI *et al.* (1993) is omitted in the WARD *et al.* (1991) analysis, presumably because it does not receive sufficient bootstrap support in the parsimony analysis. As is seen in the network, this group consists of haplotypes 18–21, and possibly of 16, 17, and 22 as well. Haplotype 17 (represented in one individual) differs from 18 (represented twice) only in substituting a C at site 324 rather than 325; this might conceivably be a baseshift reading error. (2) Group I of WARD *et al.* (1991) consists only of haplotypes 1 and 2, although there is no reason to exclude haplotypes 3 and 4 from this group; since site 362 seems hypermutable (WAKELEY 1993), haplotype 4 is likely to be derived from the neighboring haplotype 3, which in turn is separated from haplotype 2 by a single mutation at site 213. (3) In the maximum likelihood tree (obtained by a heuristic search) of WARD *et al.* (1991) we see that two links (separating haplotypes 18, 19 from the others and haplotypes 5, 6, 7 from the others) cannot be retrieved in our network. These two links constitute prime examples of what we term ghost links since they are artifactual resolutions of true polytomies: haplotype 19 is separated directly from haplotype 21 by two unique mutations, and haplotype 5 is directly derived from haplotype 8 by one unique mutation. (4) The maximum likelihood tree is one step longer than the most parsimonious trees (of length 41 steps): in the ML-tree sites 147, 278, 325, 362 together require eight steps, one more than would be necessary for the MP trees. A heuristic PAUP search delivers numerous trees of length 41. To prove that the PAUP trees are indeed most parsimonious, we apply the pruning procedure described in MATERIALS AND METHODS. We remove all haplotypes that, in view of the reduced network, could be expected to serve as intermediate nodes of some MP tree or at least be separated from such nodes by unique mutations: here one could dispense with haplotypes 2, 5, 10, 11, 12, 14, 19, 20, 25, 28, thus leaving us with 18 haplotypes, which in fact can be connected by a tree of length 41 − 7 = 34 as desired. Feeding this data into the branch-and-bound algorithm of PAUP with initial upper bound set to 34 terminates with trees of that length (in <1 hour, which seems to be faster than designing a proof by hand). A thorough visual analysis of the reduced network then provides us with 4 × 4 × 56 = 896 MP trees. (5) An earlier branching of haplotypes 6 and 7 as proposed in the published maximum likelihood tree is rather unlikely in view of the fact that the one-step network for the cluster comprising haplotypes 5–15 is starlike and nearly connected, only haplotype 7 needs two steps to connect it up most parsimoniously with haplotype 6. It is therefore plausible that the latter haplotypes are descended from haplotype 8 by a reversion at site 290.

Figure 8 shows the median network of 28 individuals of Frisian descent (data of the authors, cf. RICHARDS *et al.* 1995) between sites 90 and 365. The reduced network, highlighted in bold, is also the single most parsimonious tree. To prove that it is indeed most parsimonious, we can either apply the pruning procedure by retaining only the 11 haplotypes 9, 65, 120, 257, 258, 262, 264, 265, 266, 269, 270 for the branch-and-bound algorithm of PAUP, or give a direct proof based on counting incompatibilities: since there are four disjoint sets of pairwise incompatible sites comprising ten sites altogether (viz., {93, 224, 354}, {189, 294, 298}, {126, 304}, { 256, 270}), at least six homoplasies are necessary in a tree connecting all haplotypes. Taking into account the remaining five maximal sets of pairwise incompatible sites, it is not difficult to see that there is only one

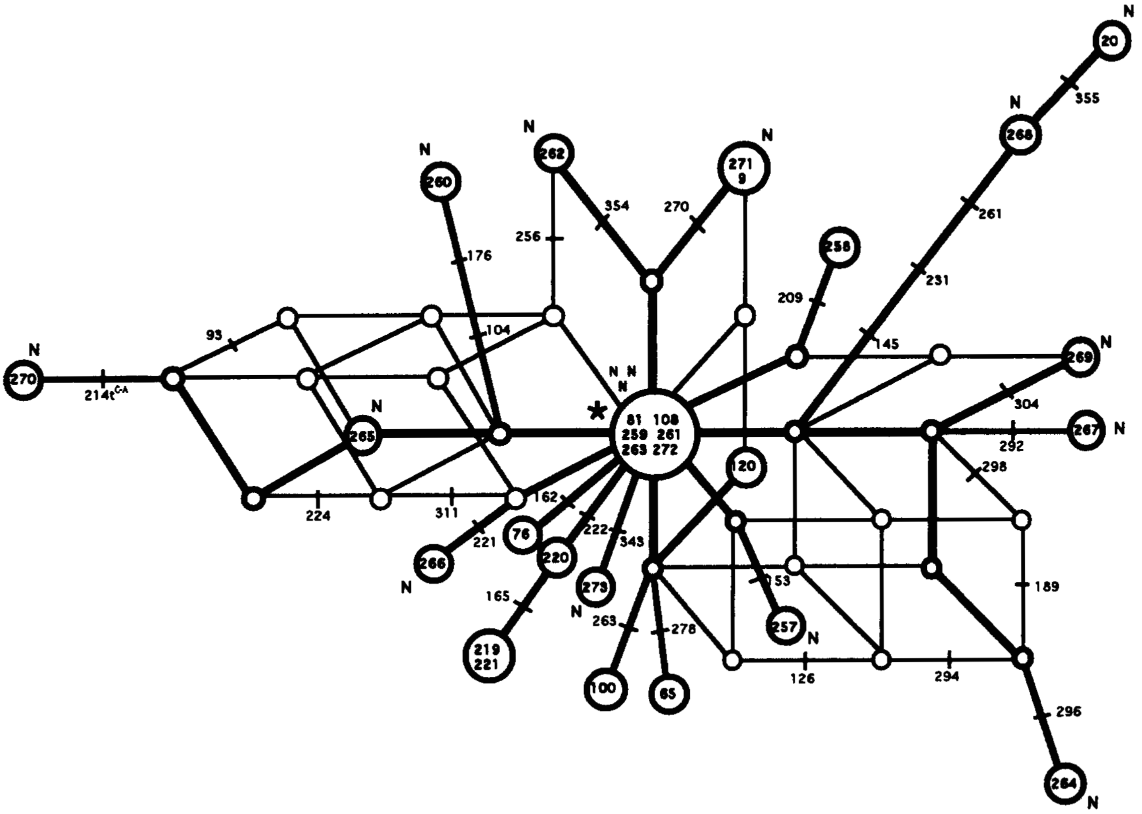

FIGURE 8.—The median network for 28 Frisians (authors' own data). N denotes individuals from North Frisian islands. The node marked by * is the consensus and simultaneously the reference sequence. The bold lines indicate the reduced network, which is the unique most parsimonious tree.

way to have no more than six sites evolved twice, namely sites 93, 189, 270, 298, 304, 354. This constraint then defines a unique tree that has realizations using only paths of the network.

The MP tree is rather star-like with little internal branching, a pattern characteristic of European data sets (DI RIENZO and WILSON 1991; PIERCY *et al.* 1993; RICHARDS *et al.* 1995; authors' unpublished data), which suggests a population-wide demographic expansion (see SHERRY *et al.* 1994). The center of the star, marked *, corresponds to the Cambridge reference sequence and is the haplotype with the highest frequency (~20%, this frequency is characteristic of European populations) and represents the consensus sequence for European haplotypes. We conclude that the Cambridge reference sequence (for this segment) is the chief ancestral haplotype of modern European populations.

DISCUSSION

In recent years a wealth of mitochondrial DNA data in the form of control region sequences as well as high-resolution RFLPs has become available for the study of human evolution (VIGILANT 1990), prehistoric migration (HORAI *et al.* 1993), and prehistoric demographic events such as sudden population expansions or severe bottlenecks (SHERRY *et al.* 1994). Dissecting population structure into plausible haplogroups and identifying the most probable paths of mitochondrial evolution would have been a valuable first step for all such studies. The methodology used to construct trees, however, did not keep abreast of the progress in obtaining mtDNA population data. Program packages, originally designed for assistance in estimating species trees, are routinely used for mtDNA data with limited success; numerous minor features and sometimes even major features of the data remain undetected this way. Parsimony methods struggle with polytomies (cf. SWOFFORD and BEGLE 1993) and overwhelm the user with an enormous number of equally parsimonious trees in the case of larger data sets. A strict consensus tree is usually too uninformative as being "hyper-polytomous" (*e.g.*, HEDGES *et al.* 1992), while selection of a single solution, without being guided by additional criteria, lacks objectivity. Unmodified distance or maximum-likelihood methods fail to give a realistic picture of the data since they return just a single tree in which polytomies are artificially resolved to some extent. This is seen quite clearly by comparing an NJ tree with an arbitrary (putative) MP tree (*e.g.*, TORRONI *et al.* 1994). Applying maximum likelihood methods that are suitable for estimating species trees to mtDNA population data would constitute a computational overkill, given that these data are often so trivial in structure (*e.g.*, quite star-like) and that relevant information (such as haplotype frequencies) for estimating coalescences is ignored anyway. Bootstrapping, usually performed as well, has little impact here on assessing reliability of estimated links. In the ideal situation the one-step network of the sampled haplotypes would be connected and free of homoplasy; then, given k sites altogether, the expected bootstrap value for a single link would equal the probability $1 - (1 - 1/k)^k$ to resample the single site supporting this link at least once; this value is ~63%. If, however, the one-step network is still connected but has just one four-cycle, reflecting a single parallel event, then any of the four links of the cycle would receive a bootstrap value of ~43% [approximating $^1/_2(1 - 1/e^2)$, provided ties are broken at random], even though one of the four spanning trees in the one-step network may clearly be favored in view of coalescent theory. Bootstrap percentages are thus generically fairly low for mtDNA population data, rendering them quite irrelevant. Thus, clusters can escape detection if one pinpoints only branches that enjoy considerable bootstrap support. Natural haplogroups cannot cogently be treated as if they were "monophyletic groups" since they may testify to different migrations and expansion times: the central haplotype of a recent cluster may be descended (after a period of isolation) from the direct ancestor of an earlier cluster.

An approach using networks rather than trees has many advantages. The median network generated by a table of binary data contains the same information as the table, yet in a much more comprehensible way. Labeled appropriately, the network can predict haplotypes, tell us where homoplasy is located, which sites mutated frequently, where a consensus sequence is, whether recombination is likely to have occurred, where to look for sequence errors, which haplogroups may be discerned, and so on. If not too large, it thus provides insight into and familiarity with the data. Since the median network harbors all most parsimonious trees for the input data, it yields a more concise picture of the data than an exhaustive list of all MP trees.

The simplest way of obtaining a network is, of course, to consider the one-step network, which alas is rarely connected. It is nearly connected though in the case of low-resolution RFLPs, which however cannot discriminate well between distant populations (cf. Excoffier *et al.* 1992). Resorting to two- or higher-step networks to connect the one-step components (as done by Excoffier and Smouse 1994) is likely to give erroneous solutions; for instance, the Frisian MP tree (Figure 8) could not be inferred by such an approach. Generally, one should not dispense with an explicit reconstruction of intermediate haplotypes, as is performed in the construction of the median network. In the median network (or the reduced network), the one-step connected components convey information about most likely paths; see the discussion on the plausible solutions in the Nuu-Chah-Nulth network (Figure 7).

If the number of haplotypes is large, the median network representing the data would become impractical due to the presence of high-dimensional hypercubes. Therefore some conspicuous parallelisms and reversals should be identified beforehand by hypothesizing hypervariability at some sites. Technically, this amounts to substituting an original character by two or more new characters. We have proposed a simple technique, executable by hand, to achieve this by employing a compatibility argument and additionally taking haplotype frequencies into account; but other strategies could serve the purpose equally well. In any case one would arrive at a revised data matrix with more characters but less homoplasy. This modified matrix in turn generates a median network, here referred to as the reduced network, that may already be quite treelike or amenable to further reduction according to criteria borrowed from coalescent theory. The reduction procedure proposed here is adjusted to the level of expected homoplasy and is justified only in the absence of recombination.

Network approaches are not only applicable to human mtDNA. As long as the data are binary or could be made such without loss of essential information, median networks or the more general distance-preserving networks (Bandelt 1994) constitute the visualization tool of choice.

We thank K. A. Crandall for helpful comments. This work was partly supported by the Wellcome Trust.

LITERATURE CITED

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. L. De Bruin, A. R. Coulson *et al.*, 1981 Sequence and organization of the human mitochondrial genome. Nature **290:** 457–465.

Bailliet, G., F. Rothhammer, F. R. Carnese, C. M. Bravi and N. O. Bianchi 1994 Founder mitochondrial haplotypes in Amerindian populations. Am. J. Hum. Genet. **55:** 27–33.

Bandelt, H.-J., 1984 Retracts of hypercubes. J. Graph Theory **8:** 501–510.

Bandelt, H.-J., 1992 Generating median graphs from Boolean matrices, pp. 305–309 in *L_1-Statistical Analysis*, edited by Y. Dodge. Elsevier, North-Holland.

Bandelt, H.-J., 1994 Phylogenetic networks. Verhandl. Naturwiss. Vereins Hamburg (NF) **34:** 51–71.

Bandelt, H.-J., and J.-P. Barthélemy, 1984 Medians in median graphs. Discrete Appl. Math. **8:** 131–142.

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

Bandelt, H.-J., and M. van de Vel, 1991 Superextensions and the depth of median graphs. J. Combin. Theory Ser. A **57:** 187–202.

Barthélemy, J.-P., 1989 From copair hypergraphs to median graphs with latent vertices. Discrete Math. **76:** 9–28.

Castelloe, J., and A. R. Templeton, 1994 Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogen. Evol. **3:** 102–113.

Crandall, K. A., and A. R. Templeton, 1993 Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. Genetics **134:** 959–969.

Di Rienzo, A., and A. C. Wilson, 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **88:** 1597–1601.

Donnelly, P., and S. Tavaré, 1986 The ages of alleles and a coalescent. Adv. Appl. Prob. **18:** 1–19.

EXCOFFIER, L., and A. LANGANEY, 1989 Origin and differentiation of human mitochondrial DNA. Am. J. Hum. Genet. **44:** 73–85.

EXCOFFIER, L., and P. E. SMOUSE, 1994 Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics **136:** 343–359.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

FARRIS, J. S., 1970 Methods for computing Wagner trees. Syst. Zool. **19:** 83–92.

FITCH, W., 1977 On the problem of discovering the most parsimonious tree. Am. Nat. **111:** 223–257.

GREENBERG, B. D., J. E. NEWBOLD and A. SUGINO, 1983 Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene **21:** 33–49.

GUÉNOCHE, A., 1986 Graphical representation of a Boolean array. Comput. Humanities **20:** 277–281.

HEDGES, S. B., K. KUMAR, K. TAMURA and M. STONEKING 1991 Human origins and analysis of mitochondrial DNA sequences. Science **255:** 737–739.

HENDY, M. D., L. R. FOULDS and D. PENNY, 1980 Proving phylogenetic trees minimal with l-clustering and set partitioning. Math. Biosci. **51:** 71–88.

HORAI, S., and K. HAYASAKA, 1990 Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am. J. Hum. Genet. **46:** 828–842.

HORAI, S., R. KONDO, Y. NAKAGAWA-HATTORI, S. HAYASHI, S. SONODA *et al.*, 1993 Peopling of the Americas, founded by four major lineages of mitochondrial DNA. Mol. Biol. Evol. **10:** 23–47.

MADDISON, D. R., 1991 African origin of human mitochondrial DNA reexamined. Syst. Zool. **40:** 355–363.

MEACHAM, C. A., and G. F. ESTABROOK, 1985 Compatibility methods in systematics. Ann. Rev. Ecol. Syst. **16:** 431–446.

MOUNTAIN, J. L., J. M. HEBERT, S. BHATTACHARYYA, P. A. UNDERHILL, C. OTTOLENGHI *et al.*, 1995 Demographic history of India and mtDNA-sequence diversity. Am. J. Hum. Genet **56:** 979–992.

NERURKAR, V. R., K.-J. SONG, N. SAITOU, R. R. MELLAND and R. YANAGIHARA, 1993 Interfamilial and intrafamilial genomic diversity and molecular phylogeny of human T-cell lymphotrophic virus type 1 from Papua New Guinea and the Solomon Islands. Virology **196:** 506–513.

PIERCY, R., K. M. SULLIVAN, N. BENSON and P. GILL, 1993 The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. Int. J. Leg. Med. **106:** 85–90.

RICHARDS, M., P. FORSTER, S. TETZNER, R. HEDGES and B. SYKES, 1995 Mitochondrial DNA and the Frisians, pp. 141–163 in *North-Western European Language Evolution,* Supplement Vol. 12, University of Odense, Denmark.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SHERRY, S. T., A. R. ROGERS, H. HARPENDING, H. SOODYALL, T. JENKINS *et al.*, 1994 Mismatch distributions of mtDNA reveal recent human population expansions. Human Biol. **66:** 761–775.

STONEKING, M., 1993 DNA and recent human evolution. Evol. Anthropol. **2:** 60–73.

SWOFFORD, D. L., 1993 *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1.1.* Computer program distributed by the Illinois Natural History Survey, Champaign, IL.

SWOFFORD, D. L., and D. P. BEGLE, 1993 *Analysis Using PAUP: Phylogenetic Parsimony, Version 3.1.1.* User's manual distributed by the Illinois Natural History Survey, Champaign, IL.

SWOFFORD, D. L., and G. J. OLSEN, 1990 Phylogeny reconstruction, pp. 411–501 in *Molecular Systematics,* edited by D. M. HILLIS and C. MORITZ. Sinauer Associates, New York.

TEMPLETON, A. R., 1993 The "Eve" hypothesis: a genetic critique and reanalysis. Am. Anthropol. **95:** 51–72.

TORRONI, A., M. T. LOTT, M. F. CABELL, Y.-S. CHEN, L. LAVERGNE *et al.,* 1994 MtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am. J. Hum. Genet. **55:** 760–776.

VAN DE VEL, M. L. J., 1993 *Theory of Convex Structures.* North-Holland, Amsterdam.

VIGILANT, L., 1990 *Control Region Sequences from African Populations and the Evolution of Human Mitochondrial DNA.* Ph.D. Thesis. University of California, Berkeley.

WAKELEY, J., 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

WARD, R. H., B. L. FRAZIER, K. DEW-JAGER, and S. PÄÄBO, 1991 Extensive mitochondrial diversity within a single Amerindian tribe. Proc. Natl. Acad. Sci. USA **88:** 8720–8724.

Communicating editor: W.-H. LI