Using Allele Frequencies and Geographic Subdivision to Reconstruct Gene Trees Within a Species: Molecular Variance Parsimony

Laurent Excoffier* and Peter E. Smouse†

**Genetics and Biometry Laboratory, Department of Anthropology and Ecology, University of Geneva, 1227 Carouge, Switzerland, and †Center for Theoretical and Applied Genetics, Cook College, Rutgers University, New Brunswick, New Jersey 08903-0231*

Manuscript received January 29, 1993
Accepted for publication September 11, 1993

ABSTRACT

We formalize the use of allele frequency and geographic information for the construction of gene trees at the intraspecific level and extend the concept of evolutionary parsimony to molecular variance parsimony. The central principle is to consider a particular gene tree as a variable to be optimized in the estimation of a given population statistic. We propose three population statistics that are related to variance components and that are explicit functions of phylogenetic information. The methodology is applied in the context of minimum spanning trees (MSTs) and human mitochondrial DNA restriction data, but could be extended to accommodate other tree-making procedures, as well as other data types. We pursue optimal trees by heuristic optimization over a search space of more than 1.29 billion MSTs. This very large number of equally parsimonious trees underlines the lack of resolution of conventional parsimony procedures. This lack of resolution is highlighted by the observation that equally parsimonious trees yield very different estimates of population genetic diversity and genetic structure, as shown by null distributions of the population statistics, obtained by evaluation of 10,000 random MSTs. We propose a non-parametric test for the similarity between any two trees, based on the distribution of a weighted coevolutionary correlation. The ability to test for tree relatedness leads to the definition of a class of solutions instead of a single solution. Members of the class share virtually all of the critical internal structure of the tree but differ in the placement of singleton branch tips.

MANY different methods have been proposed to reconstruct phylogenies from molecular data, with the details depending on the type of data (distances, character states), or possible biases (homoplasies, unequal evolutionary rates) in the evolution of the molecules (FELSENSTEIN 1988; SWOFFORD and OLSEN 1990). Several attempts have been made to incorporate reasonable biological assumptions into phylogenetic reconstruction. Dollo parsimony minimizes the number of restriction site gains over losses, concentrating the resolution on the rarer (and thus more telling) changes (TEMPLETON 1983). Generalized parsimony (SANKOFF 1983) associates costs with different types of evolutionary changes (*i.e.*, transitions, transversions, restriction site gains or losses, length variants), approximately proportional to the inverse of their probability of occurrence; the aim is to choose a phylogeny that minimizes the total evolutionary cost. LAKE's (1987) invariant method focuses on interior branch substitutions in order to reduce the effect of highly unequal evolutionary rates among distant lineages. Maximum likelihood techniques use a stochastic model of neutral evolution to weight various changes in proportion to their evolutionary information content (LI 1986; SMOUSE and LI 1987; SMOUSE *et al.* 1991; FELSENSTEIN 1988, 1992).

Traditional methods share a pair of common features: (a) they concentrate on molecular information, using one replicate observation for each molecular variant and discarding information on its population frequency and geographic location; (b) the object of the exercise is to obtain a single evolutionary reconstruction that is viewed as optimal. The classical approach is probably reasonable when the object is to create a phylogeny of duplicated genes within a species or of analogous genes among species, but it is becoming increasingly apparent that additional information on the frequencies and locations of the different variants can be useful when studying a collection of haplotypes from a single species. It is also becoming clear that more than one solution (perhaps a great many) may be acceptable for a specific problem; we need to define a class of acceptable solutions.

There are sound theoretical reasons why allele frequencies and geography might provide important information (CRANDALL and TEMPLETON 1993). There is a direct relationship between haplotype frequencies and the ages of the haplotypes (WATTERSON and GUESS 1975), information that may be useful when constructing a gene genealogy. High frequency haplotypes have probably been present in the population for a long time, having had a chance to achieve substantial copy numbers. As the vast majority of new mutants are derived from common haplotypes, we can

anticipate that rarer variants, generally representing more recent mutations, are more closely related to the common haplotypes in the extant collection than they are to other rare variants, everything else being equal. Geographic information may also tell us something useful about the relationships among haplotypes, as the immediate descendents of a new mutation are more likely to remain in the original population than to emigrate to some distant population, unless high levels of gene flow occur between those populations (SLATKIN and MADDISON 1989). As evolutionary time passes, opportunities for emigration and geographic diffusion accumulate, and the geographic frequency pattern (the population structure) of the full collection of haplotypes constitutes a long-lasting signature of the evolutionary history connecting those variants (THOMPSON *et al.* 1992).

In our search for evolutionary truth, we have traditionally concentrated on finding the single best reconstruction. Quite apart from the impossibility of evaluating a prohibitively large number of candidate trees, we have virtually no hope of designating truth with any degree of confidence, even if we could be confident of exhaustive enumeration. Given any particular phylogenetic truth, the observed data represent only one of an incredibly large number of possible outcomes, each of which (looking forward) had a vanishingly small probability of occurrence. In addition, a huge number of evolutionary truths could have given us the results we actually see, and (looking backward) none of them has large likelihood. We should not view any one solution, even the best, with any large degree of comfort. We need to begin thinking about classes of acceptable solutions that allow us to bracket our estimates and our ignorance. The dangers of focusing too closely on one particular solution are illustrated by the recent controversy over "African Eve" (MADDISON 1991; VIGILANT *et al.* 1991; HEDGES *et al.* 1992; TEMPLETON 1992), but while that case has generated much notoriety, the problem is both widespread and intrinsic to traditional practice. In any effort to define a class of credible solutions, we are led inevitably to a consideration of how those alternatives are related to each other. Is the class simply an otherwise unrelated collection of alternatives, all meeting some arbitrary criterion, or do members of the class share anything else in common? We are led to develop some measure of the similarity of different trees, choosing a metric that can also be used to compare any particular tree with the original data set.

Empirically, it is not uncommon to invoke frequency or locational information to justify a particular choice among a set of equally parsimonious trees (EXCOFFIER and LANGANEY 1989; QUATTRO, AVISE and VRIJENHOEK 1991, 1992; VIGILANT *et al.* 1991; CRANDALL and TEMPLETON 1993), but no one has explicitly included frequency or locational information into the tree evaluation process itself. There is some attention now being devoted to the range of alternative trees (FELSENSTEIN 1988; LI and GOUY 1990; MADDISON 1991; HEDGES *et al.* 1992; MADDISON, RUVOLO and SWOFFORD 1992; TEMPLETON 1992), but more work in this area is badly needed. Our intent here is to recast the phylogenetic inference problem in an overtly population genetics context. The incorporation of such information not only provides us with an additional criterion for choosing among a set of equally parsimonious solutions, but also provides us with a framework for the comparison of plausible alternative reconstructions.

We rank competing trees according to a set of criteria other than mere tree length, drawn from population analysis of variant frequency and geographic location. The overall strategy is to optimize some population statistic over the choice of tree. We first translate each competing tree into a matrix of evolutionary relationships (patristic distances) among haplotypes, one matrix per tree. Using these competing distance matrices, we then compute a set of relevant population statistics. We propose three statistics, the choice of which depends on the specific question under examination. Finally, we obtain empiric null distributions of these statistics by sampling from the restricted space of minimum spanning trees (MSTs), the equally parsimonious alternatives among which we can choose, but similar null distributions may be obtained for different tree reconstruction methods. We then introduce an heuristic procedure to find ever better trees, searching for the global optimum. Having obtained some collection of excellent candidate trees, we relate them to each other and to the defining data set, using a cophenetic correlation analog (SOKAL and ROHLF 1962). We will show that excellent trees are both highly correlated *inter se* and highly correlated with the raw data; these correlations decrease as we consider progressively suboptimal trees.

As trees are often used directly to interpret relationships among populations (CANN, STONEKING and WILSON 1987; EXCOFFIER and LANGANEY 1989; SLATKIN and MADDISON 1989; VIGILANT *et al.* 1991; MADDISON, RUVOLO and SWOFFORD 1992), we find it useful to define the similarity between two trees as a function of both tree information and haplotype population frequencies. We then apply this new methodology to a data set of human mtDNA restriction haplotypes, the same set we used earlier to illustrate the AMOVA (analysis of molecular variance) technique (EXCOFFIER, SMOUSE and QUATTRO 1992), a convenient way of cross-referencing the two sets of methods.

METHODS

Translating trees into distance matrices: The standard problem, that of creating the tree from

molecular distances, is the subject of many different phylogenetic reconstruction methods (CAVALLI-SFORZA and EDWARDS 1967; FITCH and MARGOLIASH 1967; SAITOU and NEI 1987; BULMER 1991). It is the ultimate object of this paper as well, but we begin with the converse problem, that of defining molecular distances from the tree itself. For this paper, we use spanning trees, where the operational taxonomic units (OTUs) (mtDNA haplotypes) serve both as nodes and branch tips of the tree. Spanning trees are to be distinguished from the Steiner trees more familiar in phylogenetic inference, where OTUs normally serve only as branch tips. There are some useful graph-theoretic features of spanning trees that we employ and that are not matched by analogs for Steiner trees. In addition, the spanning trees convey the flavor of the intraspecific evolution of haplotypes in a way that is more helpful for our immediate problem.

We begin by translating relationships among haplotypes, determined from a particular tree, into more mathematically tractable form. Consider any particular tree linking H haplotypes. Let $\mathbf{X}_j$ be a Boolean vector representing the jth haplotype, a string of 1's and 0's, representing the binary states of a series of positions along the molecule. This representation is most natural in the context of a restriction haplotype, but with a little care, DNA sequences can be accommodated as well. The dimension of $\mathbf{X}_j$ is set to be the number of independent and unique mutational events (m) having occurred along the tree, *i.e.*, the length of the tree. Once an arbitrary haplotype in the tree has been chosen as a reference point, any other haplotype can be encoded in vectorial form, by recording its observed differences from the reference haplotype along all m dimensions, as shown in Figure 1.

With that definition, we compute the squared evolutionary (patristic) distances between haplotypes j and k (δ^2_{jk}) along the tree as the Euclidean differences of their respective frequency vectors

$$\delta^2_{jk} = (\mathbf{x}_j - \mathbf{x}_k)' \mathbf{W} (\mathbf{x}_j - \mathbf{x}_k) \qquad (1a)$$

where $\mathbf{W}$ is an $m \times m$ square matrix of differential weights for the various mutational events. Where $\mathbf{W}$ is diagonal (assuming that mutational events are independent but providing different amounts of information), the squared patristic distance are rewritten as

$$\delta^2_{jk} = \sum_{s=1}^{m} w_s^2 (x_{sj} - x_{sk})^2 \qquad (1b)$$

where the subscript s indexes the sites. The same line of reasoning can be applied to DNA sequence data. The DNA haplotypes located on a particular phylogeny may be translated into a series of Boolean vectors if individual mutational events have been recognized, which is usually the case once a phylogeny has been imposed (see Figure 1). The w_s^2's can reflect the differential weighting of transitions versus transversions, or synonymous *vs.* code changing mutations in coding sequences. For the data in hand (EXCOFFIER, SMOUSE and QUATTRO 1992), we have shown that unequal weights do not change the outcome, so we simply use $\mathbf{W} = \mathbf{I}$, the identity matrix, weighting all mutational events equally. Our objective here is to illustrate the importance of population information for the construction of the trees, so we suppress the nuances of unequal weighting.

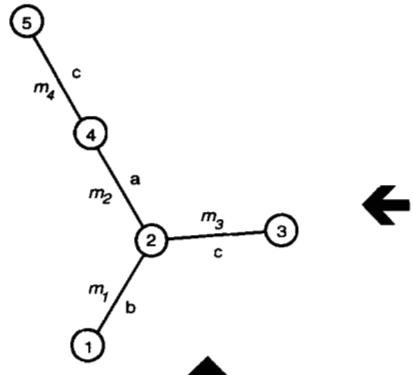
Defining population statistics as functions of the tree: In order to study population genetic structure with molecular information, we recently developed a methodology (EXCOFFIER, SMOUSE and QUATTRO 1992) that allows us to convert a Euclidean distance matrix $\mathbf{D} = \{\delta^2_{jk}\}$ into a partition of molecular variation within and among populations, as well as a set of estimated F statistic analogs (called Φ statistics) of the sort described by WRIGHT (1951, 1965; see also COCKERHAM 1969, 1973; WEIR and COCKERHAM 1984; LONG 1986; SMOUSE and LONG 1988). We will use some of these same statistics, derived from this AMOVA, as alternative criteria to be optimized over the choice of trees. One can imagine using other criteria, such as nucleotide diversity, but these population structure measures conveniently incorporate the frequency and geographic information.

We use here a simple hierarchical model of population genetic structure, with chromosomes collected in populations. We assume that the jth haplotype frequency vector from the ith population is a linear equation of the form

$$\mathbf{x}_{ij} = \mathbf{x} + \mathbf{a}_i + \mathbf{w}_{ij}. \qquad (2)$$

The vector $\mathbf{x}$ is the unknown expectation of $\mathbf{x}_{ij}$, averaged over the whole study. The effects are a for population and $\mathbf{w}$ for chromosomes within a population, assumed to be additive, random, independent, and to have the associated variance components (expected squared deviations) σ_a^2 and σ_w^2, respectively. The total molecular variance (σ^2) is the sum of variances due to differences among chromosomes within a population (σ_w^2) and those due to differences among the P populations (σ_a^2). The differences among chromosomes are assumed to arise by point mutations and not by recombination; for mitochondrial haplotypes, the assumptions are warranted. For nuclear gene haplotypes, we must always be concerned with recombination, and consequently, with the possibility of reticulated evolution. The tree making techniques presented here are standard in these respects.

The sums of squared deviations may be expressed as functions of haplotype counts (n's) and inter-hap-



Restriction haplotypes	*I*	*II*	*III*	*IV*
h_1	+	+	-	+
h_2	+	+	+	+
h_3	+	+	+	-
h_4	+	-	+	+
h_5	+	-	+	-

DNA sequences	*I*	*II*	*III*	*IV*
h_1	T	T	A	G
h_2	C	T	A	G
h_3	A	T	A	G
h_4	C	T	A	A
h_5	T	T	A	A

Boolean vectors	m_1	m_2	m_3	m_4
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	0	0	0	0
$\mathbf{x}_3$	0	0	1	0
$\mathbf{x}_4$	0	1	0	0
$\mathbf{x}_5$	0	1	0	1

FIGURE 1.—A gene tree of restriction haplotypes or DNA sequences can be encoded as a series of Boolean vectors of occurrence of mutational events from an arbitrary position on the tree. In each case, haplotype 2 was chosen as the reference and the elements of its associated vector $\mathbf{X}_2$ are set to zero. The elements of the other vectors are set to one if their associated haplotypes differ by some mutational event (m_is) from haplotype 2, and to zero otherwise. Note that the coding of haplotypes into Boolean vectors is independent of tree construction method but depends on the resulting tree topology.

lotypic distances (EXCOFFIER, SMOUSE and QUATTRO 1992)

$$\mathrm{SSD(T)} = \frac{1}{2N}\sum_{i=1}^{P}\sum_{l=1}^{P}\sum_{j=1}^{N}\sum_{k=1}^{N}\delta_{jk}^2 \qquad (3a)$$

$$= \frac{1}{2N}\sum_{i=1}^{P}\sum_{l=1}^{P}\sum_{j=1}^{K_i}\sum_{k=1}^{K_l} n_{ij}\, n_{lk}\, \delta_{jk}^2, \qquad (3b)$$

$$SSD(WP) = \sum_{i=1}^{P}\frac{1}{2N_i}\sum_{j=1}^{K_i}\sum_{k=1}^{K_i} n_{ij}\, n_{ik}\, \delta_{jk}^2, \qquad (3c)$$

and

$$\mathrm{SSD(AP)} = \mathrm{SSD(T)} - \mathrm{SSD(WP)}.$$

where K_i and K_l are the numbers of different haplotypes found in populations i and l, respectively, and n_{ij} is the copy number of haplotype j in population i.

From the AMOVA layout presented in Table 1, estimates of variance components are extracted from the total sum of squared deviations, SSD(T), and the within-population sum of squared deviations, SSD(WP),

$$\hat{\sigma}_w^2 = \frac{\mathrm{SSD(WP)}}{N-P}, \qquad (4a)$$

$$\hat{\sigma}_a^2 = \frac{(N-P)\ \mathrm{SSD(T)} - (N-1)\mathrm{SSD(WP)}}{(N-P)(P-1)_{n_0}}, \qquad (4b)$$

$$\hat{\sigma}^2 = \frac{(N-P)\mathrm{SSD(T)} - [(N-1)-(P-1)n_0]\mathrm{SSD(WP)}}{(N-P)(P-1)^{n_0}}, \qquad (4c)$$

where N is the sum of the P sample sizes and n_0 is a weighted average sample size for a single population (Table 1).

Finally, Φ_{ST}, the correlation of haplotypes within a population, relative to that of random haplotypes drawn from the total collection, is estimated as

$$\hat{\Phi}_{\mathrm{ST}} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}^2} = \frac{(N-P)\mathrm{SSD(T)} - (N-1)\mathrm{SSD(WP)}}{(N-P)\mathrm{SSD(T)} - [(N-1)-(P-1)n_0]\mathrm{SSD(WP)}}. \qquad (5)$$

We have thus defined a series of population statistics in (3), (4) and (5) that are functions of haplotype frequencies, squared patristic distances among haplotypes, and possible geographic partitioning of chromosomes into populations. It is important to realize that the tree (distance matrix) is the variable of interest for this problem.

TABLE 1

General design for analysis of molecular variance (AMOVA)

Source of Variation	d.f.	MSD	Expected MSD
Among populations	$P-1$	MSD(A)	$\sigma_w^2 + n0\sigma_a^2$
Among chromosomes within populations	$N-P$	MSD(W)	σ_w^2
Total	$N-1$		

$$n_0 = \frac{1}{P-1}\left(\sum_{i=1}^{P} N_i - \frac{\sum_i N_i^2}{\sum_i N_i}\right)$$

Defining the search space for optimum population statistics: For convenience, we consider only *spanning trees* in this study, instead of the strictly bifurcating trees used in most phylogenetic representations. Recall that OTUs may occupy internal nodes in spanning trees, whereas they occupy only terminal nodes in classical phylogenies. Spanning trees assume that the direct common ancestor of all observed haplotypes is itself present in the sample. Multifurcations are also allowed in spanning trees, so a reference haplotype may have given rise to more than two other haplotypes. These assumptions are probably valid for intra-specific studies of large samples drawn from related populations but are less tenable for inter-specific studies, where longer differentiation times lead to random loss of ancestral haplotypes. In the latter case, non-spanning trees (Steiner trees) should be used to depict molecular phylogenies.

We are looking for the tree that optimizes an associated criterion value, some particular population statistic. The only guarantee of finding it would be to conduct an exhaustive examination of all possible spanning trees. The number of spanning trees for H haplotypes is given by CAYLEY (1857)

$$S(H) = H^{H-2}, \tag{6}$$

generally a very large number. For example, with $H = 20$, $S(H) \sim 10^{23}$; there is no hope of exhaustive enumeration, and we must somehow reduce the number of candidates. We can accomplish that by applying the concept of minimum evolution (parsimony), considering only that set of spanning trees of minimum length, the set of MSTs. Efficient algorithms have been described to compute MSTs (KRUSKALL 1956; PRIM 1957), and we recall here the principle of PRIM'S algorithm, as it can be extended to enumerate all possible MSTs. (a) Start with H unconnected nodes (haplotypes). Take any node A, and find a node B whose distance from A is shortest and link node A to node B. These two connected nodes form a spanning subtree. (b) Find the unconnected node C closest to a member of the spanning subtree, say A, and connect C to A. We now have a spanning subtree linking haplotypes A, B and C. (c) Repeat step (b) until all H nodes are connected to form a spanning tree.

Possible homoplasies in the evolutionary process lead to more than one MST, but all of these MSTs are subsumed by a graph obtained by modifying steps (a) and (b) of the agglomerative Prim algorithm. Instead of connecting only one closest node to one member of the spanning subtree, we connect all unconnected nodes equally closest to that member of the spanning subtree to which C is connected, and we also connect C to all equally closest nodes of the spanning subtree. Thus, at each step of the spanning subtree expansion, we establish a list of equally minimal-distance connections. This modified Prim procedure leads to a graph defining a constrained network with closed loops. With such a graph, it is possible to associate an incidence matrix, $\mathbf{K} = \{k_{ij}\}$, called a Kirchoff matrix (*i.e.*, see GIBBONS 1985), from which we can compute the number of spanning trees. The elements of $\mathbf{K}$ are defined as $k_{ij} = -1$, if nodes i and j ($i \neq j$) are connected and 0 otherwise, k_{ii} = the number of other nodes to which node i is connected.

The number of different spanning trees, $T(\mathbf{K})$, contained in this graph (network) may then be derived from the Kirchoff matrix as

$$T(\mathbf{K}) = \det \mathbf{K}_{rr}, \tag{7}$$

where $\mathbf{K}_{rr}$ is an $(H\text{-}1) \times (H\text{-}1)$ matrix derived from $\mathbf{K}$ by deleting any one row r and its associated column r (GIBBONS 1985, p. 50). Note that $T(\mathbf{K})$ is an upper bound for the number of MSTs, because not all of these trees are minimum spanning if the connection lengths in any loop of the network are not all equal. On the other hand, if all permitted links are of the same length, then $T(\mathbf{K})$ is the number of different minimal spanning trees. In either case, $T(\mathbf{K})$ is generally much less than the total number of spanning trees, $S(H)$. We illustrate with a simple example of five haplotypes, whose distance matrix is

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 1 & 2 \\ 1 & 2 & 0 & 1 & 1 \\ 2 & 1 & 1 & 0 & 1 \\ 2 & 2 & 1 & 1 & 0 \end{pmatrix}$$

The constrained graph found with the modified PRIM procedure (Figure 2a) has all links equal to unity in length, and has the associated KIRCHOFF matrix

$$\mathbf{K} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & 0 & -1 & 0 \\ -1 & 0 & 3 & -1 & -1 \\ 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

The number of minimum spanning trees is given by Equation 7 as $T(\mathbf{K}) = 11$, and these trees are shown in Figure 2.

Heuristic search for minimum spanning trees: If the number of spanning trees given by (7) is simply too large to review (say if $T(\mathbf{K}) > 100{,}000$), we can adopt a heuristic approach to optimizing our population criterion over the choice of spanning trees. The principle of the heuristic search is to pursue an optimum solution with a partial exploration of the solution space: We start from an arbitrarily chosen tree and modify its topology by "subtree pruning and regrafting" (SWOFFORD and OLSEN 1990). We evaluate the resulting population statistic; if the population criterion exceeds a predetermined threshold level, the tree topology is further modified; otherwise we go one

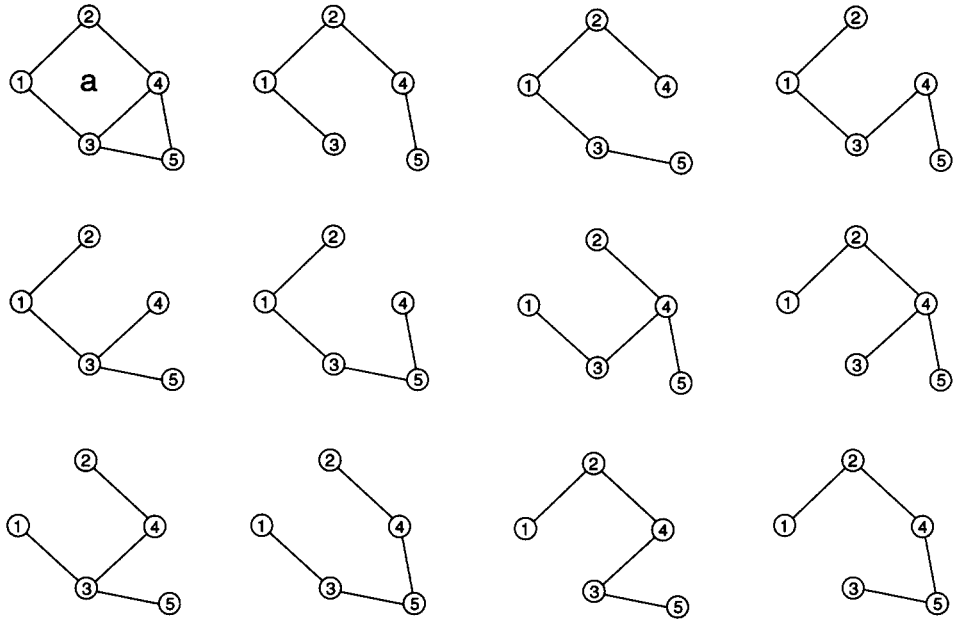


FIGURE 2.—(a) A graph (network) connecting 5 nodes. The other 11 graphs are the complete set of different spanning trees that may be derived from graph 2a.

step backward and try another subtree pruning and regrafting of the former tree. This recursive procedure stops when all possible modifications of the initial tree have been evaluated. The set of all possible topological modifications is given by the alternative connection list obtained during the construction of the minimum spanning network by the modified PRIM procedure. The obvious strategy is to continue to explore only those modified trees that improve the population criterion. With this greedy algorithm, the criterion will always improve, and the algorithm will effectively climb the nearest peak of the solution space. The solution space might be multi-peaked, of course, so there is no guarantee that the final solution will be the global optimum. Both the speed and greed of the algorithm depend on the threshold value and the possibility of temporarily tolerating non-improvement of the chosen criterion. To reduce the probability of finding a sub-optimal tree, one could either relax the threshold criterion somewhat, using a simulated annealing approach, or repeat the search several times, starting from different initial spanning trees [see SWOFFORD and OLSEN (1990) for a full discussion of greedy algorithms].

Generating random minimum spanning trees: To obtain the approximate null distribution of the population criterion over the MST-space, we can generate a large sample of random MSTs via PRIM's procedure, applied to a series of randomly ordered sets of haplotypes. By definition, each MST will have exactly the same total length, but the connections will differ from tree to tree. The procedure for generating random MSTs simply assumes that all connections between equidistant haplotypes have the same probability, a more constrained but otherwise similar procedure to that used by MADDISON and SLATKIN (1991) to generate random joining trees. There is no guarantee of finding the global optimum tree in this fashion, or even a very good tree, but it does provide a convenient way to explore the MST space widely. In principle, one might use a combination of random sampling and heuristic optimization techniques to search for alternative peaks. Our purpose here is more to explore the MST space than it is to find the global optimum, so we shall not belabor the point further.

Computing and testing correlation among trees: At the intraspecific level, a molecular tree is rarely the real object of the exercise; rather, it is used to address a specific question at the population or species level (AVISE 1989; AVISE *et al.* 1987). Since there are many competing trees, each yielding (potentially) different biological inference, it would be useful if we could devise some way to characterize the relationships among these competing trees. Any tree may be translated into a unique matrix of squared patristic distances, so an obvious way to compare any two trees is to compute the correlation between their derivative distance matrices, an analogue of the *cophenetic correlation coefficient* between the raw phenetic distance matrix and that of a particular tree (SOKAL and ROHLF 1962), a construct we shall here term a *coevolutionary correlation*. The dimension of both matrices is N, which is often large enough to be awkward for this sort of matrix manipulation, but we can finesse that problem by making use of the fact that the number of haplotypes (H) is somewhat smaller than the number of individuals (N). We only need the sums of squared deviations within each of the two matrices and the sums of cross products between the matrices

to compute the correlation between them. If we denote the distances of the first matrix by δ^2_{jk} and those of the second matrix by ϵ^2_{jk}, then we can write

$$SS(\mathbf{X}) = \sum_{j=1}^{N}\sum_{k=1}^{j-1}\left(\delta^2_{jk} - \bar{\delta}^2\right)^2 \tag{8a}$$

$$= \sum_{j=1}^{H}\sum_{k<j} n_j n_k \left(\delta^2_{jk} - \bar{\delta}^2\right)^2 + \sum_{j=1}^{H}\frac{(n_j-1)n_j}{2}\left(\delta^2_{jj} - \bar{\delta}^2\right)^2$$

$$SS(\mathbf{Y}) = \sum_{j=1}^{N}\sum_{k=1}^{j-1}\left(\epsilon^2_{jk} - \bar{\epsilon}^2\right)^2 = \sum_{j=1}^{H}\sum_{k<j} n_j n_k \left(\epsilon^2_{jk} - \bar{\epsilon}^2\right)^2 \tag{8b}$$

$$+ \sum_{j=1}^{H}\frac{(n_j-1)n_j}{2}\left(\epsilon^2_{jj} - \bar{\epsilon}^2\right)^2$$

$$SP(\mathbf{X},\mathbf{Y}) = \sum_{j=1}^{N}\sum_{k=1}^{j-1}(\delta^2_{jk} - \bar{\delta}^2)(\epsilon^2_{jk} - \bar{\epsilon}^2)$$

$$= \sum_{j=1}^{H}\sum_{k<j} n_j n_k (\delta^2_{jk} - \bar{\delta}^2)(\epsilon^2_{jk} - \bar{\epsilon}^2) + \sum_{j=1}^{H}\frac{n_j(n_j-1)}{2}(\delta^2_{jj} - \bar{\delta}^2)(\epsilon^2_{jj} - \bar{\epsilon}^2)$$

where, n_j and n_k are the observed copy numbers of haplotypes j and k, respectively. The weighted coevolutionary correlation between the two trees is then computed as

$$\hat{\rho}(\mathbf{X},\mathbf{Y}) = \frac{SP(\mathbf{X},\mathbf{Y})}{\sqrt{SS(\mathbf{X})\,SS(\mathbf{Y})}}. \tag{9}$$

The fit of any particular tree to the original data can be similarly obtained by measuring the cophenetic correlation coefficient $\hat{\rho}(\mathbf{O},\mathbf{X})$ between the phenetic distance matrix (no tree assumed) and the patristic distance matrix from that particular tree.

The significance of the correlation between two trees cannot be tested using conventional approaches; there are at least three reasons. (1) The $N(N\text{-}1)/2$ pairwise distances among the N individuals are not independent, as a consequence of both sampling realities and a common evolutionary history. (2) Each tree connects H OTUs with H-1 linearly independent branches, which can be considered as H-1 linearly independent contrasts. It follows that there cannot be more than H-1 uncorrelated trees. (3) The null hypothesis cannot be the absence of correlation, because all trees are autocorrelated, due in part to the fact that the patristic distance between identical haplotypes is always zero and that between different haplotypes is never zero.

We must therefore test whether two trees are more correlated than the average background level. After having computed $\hat{\rho}(\mathbf{X},\mathbf{Y})$, we generate a large number of random MSTs (as described above) and compute the weighted coevolutionary correlations between one of our former distance matrices (say $\mathbf{X}$) and a matrix $\mathbf{Z}$ different for each random tree, $\hat{\rho}(\mathbf{X},\mathbf{Z})$. An estimate of the probability of $\hat{\rho}(\mathbf{X},\mathbf{Y}) > \hat{\rho}(\mathbf{X},\mathbf{Z})$ is then obtained by enumeration of the null (random) distribution of $\hat{\rho}(\mathbf{X},\mathbf{Z})$. The weighted correlation measure may be viewed as a similarity index, so we can test whether two trees are significantly similar or different at a certain confidence level, assuming that we have randomly sampled within the whole universe of solutions. We have restricted attention to the set of all MSTs, and are therefore testing whether a given pair of MSTs is more similar than a random pair of MSTs.

RESULTS FROM AN ANALYSIS OF HUMAN mtDNA RESTRICTION HAPLOTYPES

We illustrate these new developments with a human mtDNA restriction-site data set described in EXCOFFIER, SMOUSE and QUATTRO (1992). The restriction-site patterns of the 56 haplotypes encountered among the 672 individuals from 10 European, Asian, African, and Amerindian populations are shown in Table 5 (see APPENDIX). A total of 62 restriction-sites were examined, 34 being polymorphic. Population haplotype frequencies can be found in Table 6 (see APPENDIX). An AMOVA of the haplotype data revealed significant differences among populations, differences that were more pronounced with a distance metric that accounted for mutational divergence among haplotypes (EXCOFFIER, SMOUSE and QUATTRO 1992). For that earlier study, we employed a single minimum spanning tree (referred hereafter as the "published tree") to compute patristic distances (EXCOFFIER, SMOUSE and QUATTRO 1992). Here, we apply our tree evaluation procedures to that same data set, describing differences between optimum spanning trees and the published tree.

The number of equally parsimonious trees exceeds one billion: We present a modified PRIM graph (Figure 3) that relates the 56 sampled haplotypes shown in Table 5 (see APPENDIX). All haplotypes may be related to each other by single restriction-site changes, except for haplotypes 29 and 11; 29 can only be connected to 9 by means of an intermediate (but missing) haplotype; 11 can be connected to 21, 23 or 46, but also requires an intermediate (but missing) haplotype in each case. The graph thus requires two (missing) intermediates to connect all haplotypes. The missing haplotypes have never been found in the more than 60 populations that have now been sampled with the same battery of enzymes (MERRIWETHER *et al.* 1991). A set of minimum spanning trees is thus a very reasonable depiction of the situation. The number of

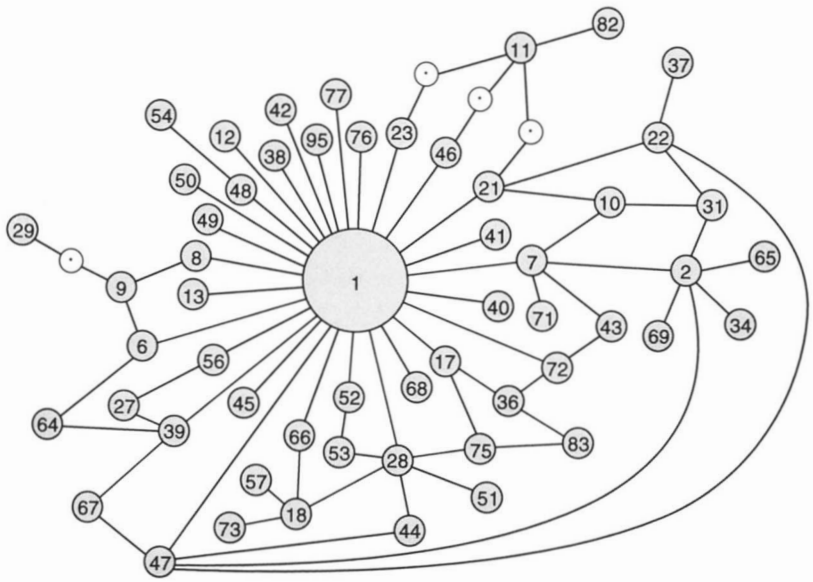


FIGURE 3.—A graph (network) relating the 56 haplotypes defined in Table 5 (see APPENDIX) by minimum connection length. Any two haplotypes are connected if they differ by a single restriction-site difference.

spanning trees given by Equation 7 is in excess of 4.4 billion (4,402,434,912). Not all these trees are MSTs, because a tree with haplotype 11 linked to two missing haplotypes has a total length of 58, compared with a length of 57 for an MST. It is nevertheless possible to compute the total number of MSTs as three times the number of spanning subtrees derived from a graph where haplotype 11 and 82 have been removed. All spanning subtrees derived from such a reduced graph would be minimum spanning, and there are three possible minimum length connections for haplotype 11 on each of these spanning subtrees. Therefore the total number of MSTs is in excess of 1.29 billion ($3 \times 431,648,856$), still far too many to permit an exhaustive search. One is forced to the conclusion that mutational parsimony is utterly useless as a sole criterion of excellence or of the best tree.

Null distributions of criteria over the MST space: We show the approximate null distributions of SSD(WP), SSD(T) and Φ_{ST} values over the MST space in Figure 4, plotted against the weighted cophenetic correlations, $\hat{\rho}(\mathbf{O},\mathbf{X})$, between patristic and phenetic distance matrices. The null distributions for σ_a^2 and σ_w^2 are not shown, inasmuch as they are mere combinations of and are highly correlated with the SSDs. The probability distributions were obtained from 10,000 random MSTs (of the 1.29 billion possible), among which 1,000 were randomly chosen and their weighted cophenetic correlations plotted. The distributions of the population criteria are shown in Figure 4 and are highly irregular, with a series of peaks and valleys. We have drawn four series of 10,000 random MSTs (results not shown), which all show the same major peaks at exactly the same positions, suggesting that these peaks reflect the major features of the graph in Figure 3. The distributions of the SSD criteria, with a large proportion of MSTs close to the optimum, suggests that a large number of alternative trees, all about equally good, differ by a topological changes involving low-frequency haplotypes that do not fundamentally alter the "essential pattern" of the tree.

The null distributions for SSD(T) and SSD(WP) present similar patterns (Figure 4, a and b), and the two criteria are strongly correlated across trees ($r = 0.988$, Table 2). The modal value is very close to the optimal value; the null distributions are skewed to the right and the weighted cophenetic correlations, $\hat{\rho}(\mathbf{O},\mathbf{X})$, are strongly and negatively correlated with SSD(WP) ($r = -0.886$) and with SSD(T) ($r = -0.879$), suggesting that very good MSTs conform better to the raw data than do sub-optimal trees (Table 2). This follows immediately from the fact that both cophenetic correlations and SSD criteria are derived directly from the underlying distance matrices among the H haplotypes (CAVALLI-SFORZA and EDWARDS 1967; FITCH and MARGOLIASH 1967; BULMER 1991).

The null distribution of Φ_{ST} (Figure 4c) is clearly different from those of the SSD criteria; three well differentiated peaks with high cophenetic correlations are distinguishable. The first peak is centered around 0.15; the second and third peaks are centered at about 0.185 and 0.225, respectively. The value computed on the original phenetic distance matrix of restriction site differences is $\Phi_{ST} = 0.227$, embedded well within the third peak. At least some of the trees in each peak have elevated cophenetic correlations, but there is a low overall correlation (although significant) between cophenetic correlation and Φ_{ST} ($r = -0.291$). The Φ_{ST} criterion is not highly correlated with the SSD criteria either [$r = 0.237$ for SSD(WP) and $r = 0.376$ for SSD(T)]. All of these observations suggest that Φ_{ST} may not be suitable as an optimization criterion. We

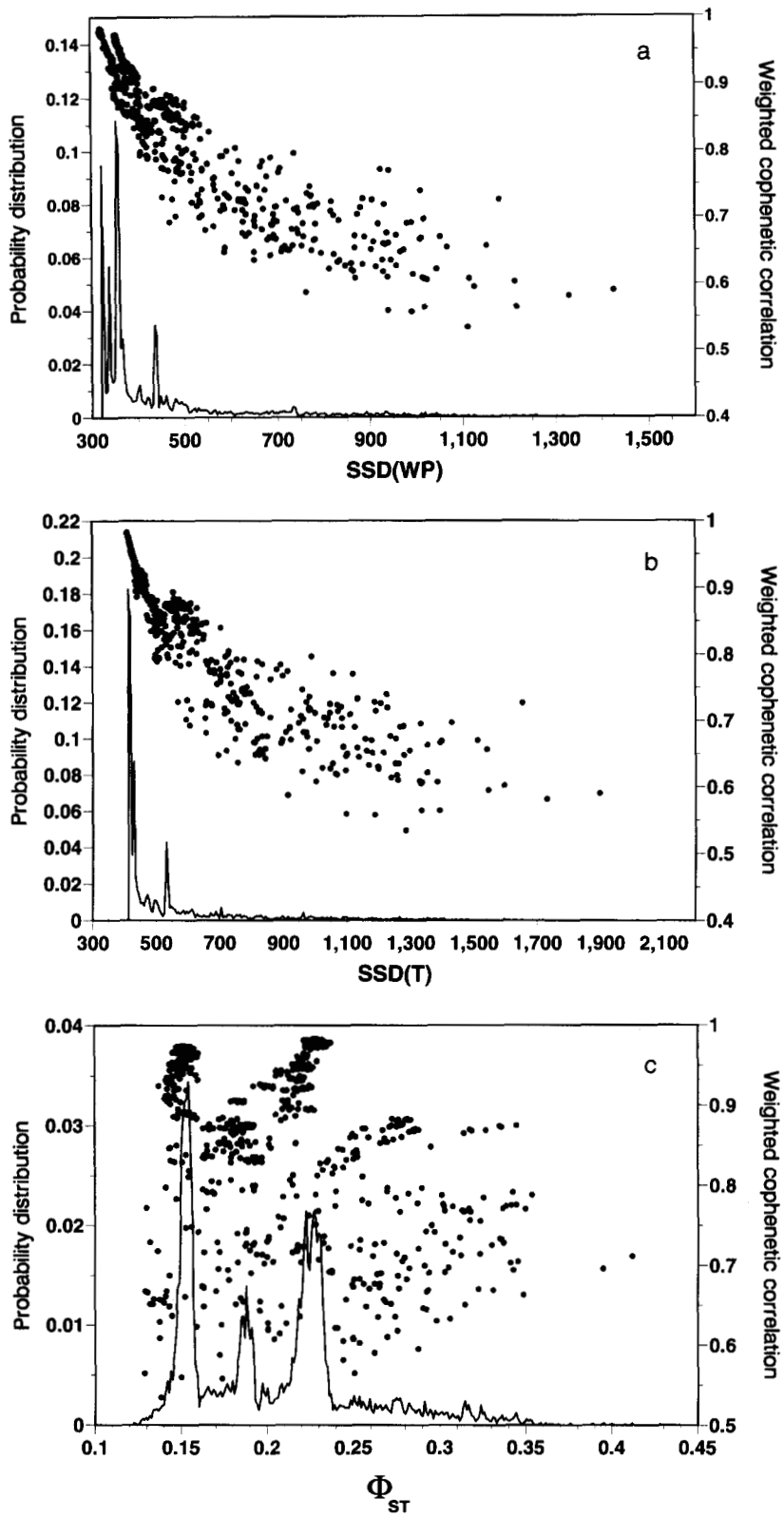


FIGURE 4.—Approximate null distribution of three population criteria. (a) SSD(WP), (b) SSD(T), (c) Φ_{ST}. The values of the probability distributions were computed on 10,000 random MSTs and are shown as continuous lines. The weighted cophenetic correlation of 1,000 random MST are also plotted against their population criteria.

would suggest that Φ_{ST} be used as characterizing output, rather than defining input; as such, it will prove useful in relating variance components to differentiation times, as shown later.

The null distributions of different population statistics are described numerically in Table 3. A striking feature of these null distributions is their wide range. They vary between 319.279 and 1540.396 for SSD(WP) and between 410.914 and 2028.679 for SSD(T), representing almost a five-fold range for the inferred variances. Equally parsimonious trees can thus yield profoundly different estimates of popula-

TABLE 2

Correlations among population statistics over the minimum spanning tree space

Population statistics	SSD(WP)	SSD(T)	σ^2	Φ_{ST}
SSD(T)	0.9875			
σ^2	0.9849	0.9999		
Φ_{ST}	0.2369	0.3758	0.3887	
$\hat{\rho}$(O,X)	−0.8858	−0.8790	−0.8772	−0.2907

All correlation coefficients are significantly different from zero at level $P = 0.001$.

tion genetic diversity and, collaterally, divergent estimates of population mean nucleotide diversity and differentiation times. The lower ($P < 0.005$) confidence limits for SSD(WP) and SSD(T), among randomly constructed MSTs, are 320.341 and 411.333, respectively (Table 3). The values for the MST published in EXCOFFIER, SMOUSE and QUATTRO (1992; Figure 5) were 320.25 and 411.02 for SSD(WP) and SSD(T), respectively. That earlier tree (intuitively based on frequency and geographic information but without rigorous quantification) is superior to the vast majority of MSTs that emerge from a random sampling.

Optimal trees present a conserved structure: Using SSD(WP) and SSD(T) criteria for optimization, we next examine the topological structures of the 100 best trees, recording the frequencies of each possible connection between haplotypes. There are 28 unambiguous connections that are common to all 1.29 billion MSTs, connections that can only be made in one way for an MST. In addition, a clear common structure emerges from the 100 best (1%) trees, a series of additional (and nonobligatory) connections that are nevertheless inevitably present. Haplotype 2 is always connected to haplotype 7 in the 100 best SSD(T) and SSD(WP) trees; haplotype 10 is always connected to haplotype 7 in the 100 best SSD(WP) trees. Moreover, some permissable connections are never used for these best trees: these are connections 2–47 and 10–21 for the 100 best SSD(WP) trees, and 2–47, 8–9, and 22–31 for the 100 best SSD(T) trees. Other connections (Figure 3) appear between 3 and 97 times; their presence or absence has less impact on the SSD criteria.

Correlations among different optimal trees: There is much to be learned from a comparison of our previously published tree with these 100 best trees. The published tree (EXCOFFIER, SMOUSE and QUATTRO 1992; Figure 3) is presented in Figure 5a; we present that which minimizes SSD(WP) in Figure 5b. This latter tree, which may be considered as having considerable geographic structure, differs from that in Figure 5a at five points; haplotype 53 is connected to haplotype 28 in Figure 5b, instead of to haplotype 52 in Figure 5a; haplotype 83 is connected to 75 instead of 36; haplotype 31 to 22 instead of 2; haplotype 44 to 47 instead of 28; and haplotype 22 to 47 instead of 21. These differences mainly involve haplotypes that are located at the very tips of the tree, found only once in the total sample. The exception is the connection of haplotype 22 to 47 instead of 21. The MST that minimizes SSD(T) is shown in Figure 5c. There are six topological differences between this tree and the published tree in Figure 5a (involving haplotypes 9, 36, 53, 64, 67 and 83) and seven differences between this tree and the minimum SSD(WP) tree of Figure 5a (involving haplotypes 9, 22, 31, 36, 44, 64, and 67). These topological differences also involve only the rare haplotypes located at the branch tips.

The similarities among these three MSTs are better understood when comparing them with a randomly chosen MST having large sums of square deviations, an example of which is shown in Figure 5d (SSD(WP) = 1009.901; SSD(T) = 1398.905). This latter tree shows major topological differences from the trees in Figure 5, a-c: it has much longer branch lengths, resulting in larger patristic distances among haplotypes and much larger SSDs. Such topological outcomes are typical of all the suboptimal trees we have examined, and it is important to recall that these suboptimal trees are all MSTs and molecularly parsimonious. Some parsimonious trees are substantially better than others.

The weighted correlations between the 6 pairs of trees from Figure 5 are reported in Table 4. The significance of these correlations was assessed as described earlier, using 1,000 randomly selected MSTs. The only significant correlation ($P = 0.016$) is that

TABLE 3

Random minimum spanning tree distribution statistics

Population statistics	Minimum	Maximum	Mean (SD)	Percentile limits 0.1%	0.5%	1%	5%
SSD(WP)	319.279	1540.396	445.847 (479.396)	319.729	320.341	320.677	321.856
SSD(T)	410.914	2028.679	558.133 (606.238)	411.104	411.333	411.512	412.339
σ^2	0.629	3.116	0.853 (0.927)	0.629	0.630	0.630	0.631
Φ_{ST}	0.122	0.429	0.202 (0.208)	0.128	0.137	0.137	0.147

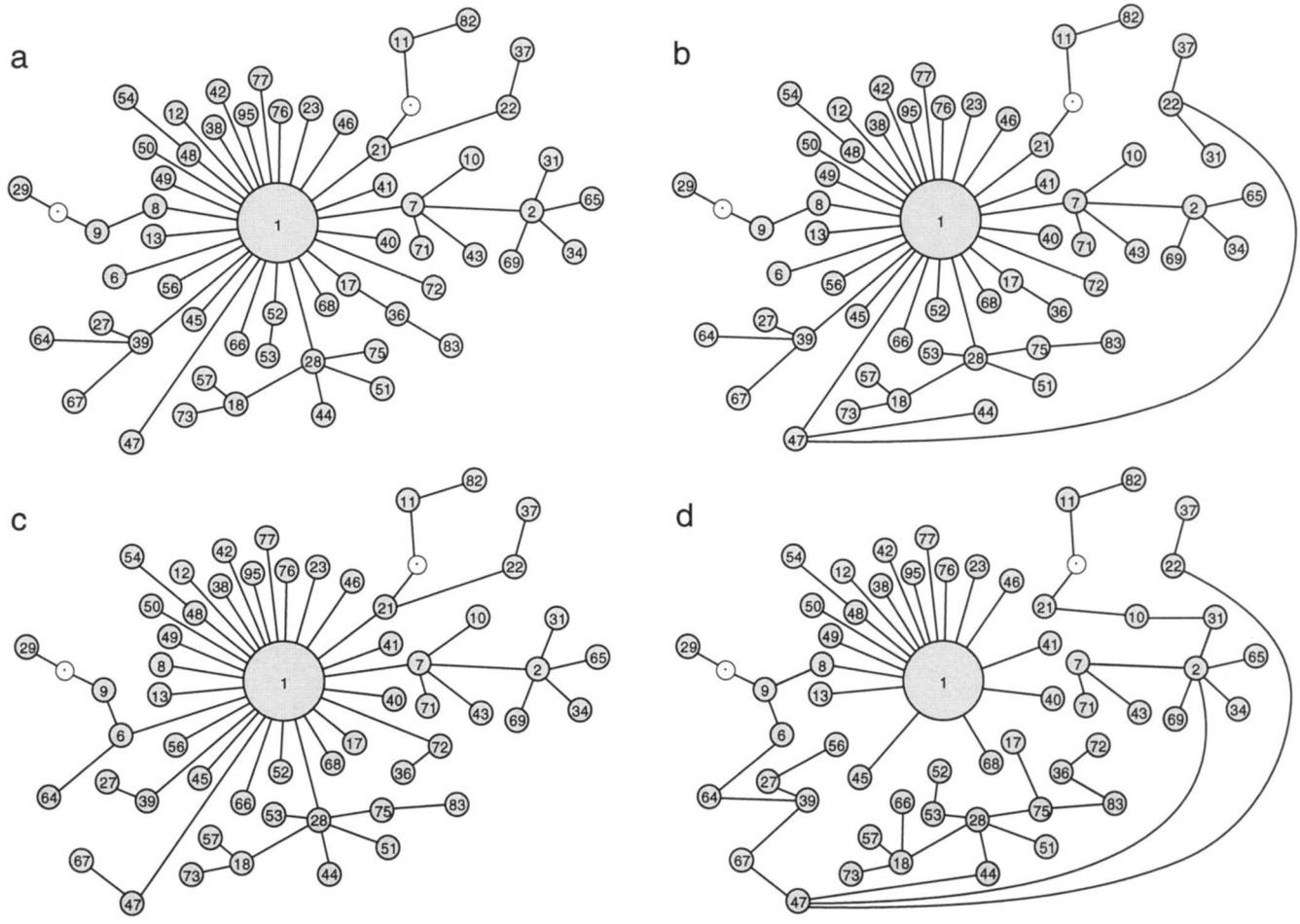


FIGURE 5.—Four parsimonious trees having the same minimum length but implying different molecular variances at the population level. (a) MST previously published in EXCOFFIER, SMOUSE and QUATTRO (1992) (SSD(T) = 411.82; SSD(WP) = 320.25). (b) MST minimizing SSD(WP) (319.203) and σ_w^2 (0.482). (c) MST minimizing SSD(T) (410.842) and σ^2 (0.629). (d) Random MST with large associated sums of square deviations values (SSD(WP) = 1009.901 and SSD(T) = 1398.905), leading to considerably inflated molecular variances (σ_w^2 = 1.526, σ^2 = 2.159).

TABLE 4

Weighted coevolutionary correlations among trees shown in Figure 5

Trees	Published tree (Figure 5a)	Optimum for SSD(WP) (Figure 5b)	Optimum for SSD(T) (Figure 5c)	Random MST (Figure 5d)
Published tree (Figure 5a)	—	0.150	0.016	0.299
Optimum for SSD(WP) (Figure 5b)	0.9799	—	0.156	0.495
Optimum for SSD(T) (Figure 5c)	0.9964	0.9767	—	0.249
Random MST (Figure 5d)	0.6406	0.6310	0.6439	—

Weighted correlations are shown below the diagonal, and the probabilities of getting a larger correlation by chance alone are shown above.

between the published tree and the tree minimizing SSD(T). Surprisingly, although their topological differences are small, the correlation between the optimum SSD(T) and SSD(WP) trees and that between the published tree and the optimum SSD(WP) are not significant. The lack of statistical correlation derives from the fact that the mean autocorrelation level among all MSTs is extremely high, due to the shared connections within the entire set. Correlations between the suboptimal MST (Figure 5d) and the other trees are lower and less significant, reflecting profound structural differences.

In summary, it would appear that while there is an enormous number of (equally) maximum parsimony trees, the situation is far from hopeless. If we employ some other criterion, in addition to molecular parsimony, there is a great deal to choose among these alternative trees. We have shown, using frequency and geographic information, in conjunction with an AMOVA, that there is a tremendous range of per-

formance among equally parsimonious trees. There nevertheless remain many excellent candidate trees among which to choose, still too many to evaluate exhaustively. If they shared nothing in common, we would find ourselves in an awkward position, but we have shown that the excellent candidates share all of the *critical internal structure* of the tree, differing only in the branch tip placement of rare haplotypes. Moreover, all members of the class have high cophenetic correlations with the phenetic distance matrix, and represent minimal distortions of the data in the process of tree construction. All excellent MSTs are members of a single class of solutions that conforms closely to the raw data and that protects this internal structure. Interestingly, our earlier, hand-drawn MST, with some careful attention devoted to frequency and geographic information, was a superior member of the class, suggesting that such a construct might easily identify that critical internal structure and provide a departure point for a heuristic optimization search.

DISCUSSION

Molecular variance parsimony: We have presented here a new technique for reconstructing intraspecific molecular trees that makes use of sampled haplotype frequencies and geographic information. It extends the notion of conventional parsimony to "molecular variance parsimony" in the sense that we not only minimize the total number of mutational events having occurred in the past (the length of the tree), but also the population molecular variance (a function of both molecular differences and allele frequencies). Although specifically applied here to minimum spanning trees, our methodology could be applied to other phylogenetic reconstruction methods, providing one or more additional criteria to be evaluated.

The limits of conventional parsimony: The distribution of different population statistics (Figure 4) for equally parsimonious trees shows that trees having exactly the same mutational length may have very different properties and may sometimes fit the original data poorly. It follows that the parsimony is not adequate as a sole criterion for choosing an optimum gene tree at the intraspecific level, as recently shown in the context of mtDNA sequences (HEDGES *et al.* 1992; MADDISON, RUVOLO and SWOFFORD 1992; TEMPLETON 1992; VIGILANT *et al.* 1991), where the origin of modern humans was inferred from an arbitrarily chosen phylogenetic tree. Two different problems arise with the parsimony criterion in the presence of homoplasy: (1) it is often not very discriminating, as attested by the very large number (1.29 billion) of equally parsimonious MSTs, and (2) it does not lend itself well to the definition of confidence intervals around the most parsimonious state or to significance testing (FELSENSTEIN 1983, 1988). These two problems are circumvented by the present method, which imposes one or more additional criteria, in an attempt to define a class of excellent solutions and then to characterize that class.

The use of allele frequencies in phylogeny reconstruction raises the problem of the sample size in molecular population studies. Common belief is that sample size is not very important in molecular studies, because it does not affect the coefficient of variation of nucleotide diversity estimates as much as the number of nucleotides surveyed or the number of loci (NEI 1987). It follows from this view that one should usually prefer to study many loci on a few individuals, rather than many individuals for a few loci. On the other hand, inasmuch as nucleotide diversity should be computed along a given tree, this line of reasoning is only correct if one assumes that the true tree is known. For the example given here, the molecular variance can change over a fivefold range (Table 3), depending on which of the 1.29 billion MSTs is considered, among which there is no choice in the absence of allele frequency or population structure information. That fact suggests the need for reliable allele frequency estimates for molecular population studies, which can only be obtained with large sample sizes.

One solution *vs.* a class of solutions: Although we may drastically reduce the number of plausible trees, relative to the total number of MST trees, our results suggest that it would be unreasonable to settle firmly on any single "best" tree. There are a great many excellent trees from which to choose, and the choice depends to some extent on which statistic is selected for optimization. When population structure is reasonably well established, we might prefer a phylogeny that minimizes SSD(WP) and σ_w^2, leading to minimum overall differences among haplotypes from the same population. This would be a tree with the strongest geographic coherence and with the smallest average coalescence time of chromosomes within each population. A tree maximizing σ_a^2 [or SSD(AP) = SSD(Total) − SSD(WP)] would maximize divergence among populations, possibly proving useful in the context of population discriminant analysis. On the other hand, we might prefer simply to minimize σ^2 or SSD(T) without regard to population structure. This would be a particularly logical choice if the point were to provide a population-structure-neutral tree, along which one could simply measure population divergence. Our example suggests that Φ_{ST} may not be a useful criterion to optimize. Although a large Φ_{ST} value would indicate a high level of population differentiation, trees associated with the highest Φ_{ST} values sometimes distort the original data. In other words, an optimization of population structure (Φ_{ST}) could lead to an appearance of population structure in a globally random mating population. In principle, any

other statistic based on patristic distances, like nucleotide diversity (NEI and TAJIMA 1983), could be optimized as well, and lead to its own optimum. Moreover, we do not have any evidence that there is a single optimum for each statistics. In the case of SSD(T), we were able to find at least two trees having exactly the same value, but with a single topological difference. We would therefore recommend repeating the heuristic optimization procedure from several different starting points, if the computing time is not prohibitive.

Although the point of this study is more to examine the impact of the choice of a particular phylogeny on the estimation of a population parameter than to come up with a single resulting phylogeny, we are led to ask ourselves how close we are to the "true tree"? As is the case for every phylogenetic inference method, we have no formal proof that our procedure will always lead to the true tree; quite apart from the difficulty of discovering truth, the method is no better than the underlying assumptions. For instance, parsimony assumes that the number of mutations during the evolutionary process is minimized, whereas maximum likelihood methods assign probabilities to each mutation type, favoring the evolutionary process made up of a series of most likely mutational events. One can imagine real evolutionary processes departing from either set of defining assumptions, but one generally views those assumptions as reasonable. Our methods impose additional assumptions; frequent haplotypes should be as close as possible on the phylogeny, and if molecular variance within population is minimized, haplotypes found in the same population should also be closely related. These additional assumptions make intuitive sense and have been shown to be valid over a coalescent differentiation process (CRANDALL and TEMPLETON 1993). The frequency assumption simply states that there has been a single molecular differentiation center, while the within-population minimization approach implies that the migration rate among populations has been reduced, relative to panmictic expectation.

Choosing a single optimal tree is not entirely satisfying, as topologically similar trees furnish similar solutions to the underlying population problem. Some topological differences may have less influence on the population statistics than do others, which is why we have proposed computation of a similarity index (a weighted coevolutionary correlation coefficient) between any two trees, not only on the basis of the trees themselves, but also on population information. The significance of such a correlation is obtained through its distribution over the solution space, permitting definition of a class of solutions based on a statistical criterion of those trees having the highest 1% or 5% coevolutionary correlation coefficients with the optimum tree, and we have seen that these trees clearly share substantial (and all of the critical) topological structure. The solution class thus represents a confidence statement about the inference neighborhood of the optimum solution. The size of this class depends not only on the tree topology and the population genetic structure, but also on the total number of available trees in the solution space. The class limits would certainly change if we consider non-parsimonious trees (non-MSTs). There could exist some trees of greater total length that would yield better population statistics than do some of the poorer MSTs. We have arbitrarily chosen not to consider non-MSTs in the present study, partly for simplicity but also because the parsimony criterion is reasonable qualifying criterion. It is possible to be more general, of course, optimizing strictly bifurcating trees instead, without any *a priori* ability to specify the minimum length.

Minimizing coalescence times and migration events: By minimizing molecular variance, we also minimize population overall differentiation times, because the molecular variance components can be explicitely related to mean coalescent times. This is a consequence of the fact that Φ_{ST} is the expected ratio of mean coalescence times ($\bar{t}$'s), as shown by SLATKIN (1993)

$$\Phi_{ST} = 1 - \frac{\bar{t}_0}{\bar{t}_1}, \qquad (10)$$

where $\bar{t}_0$ is the average coalescence time of two genes drawn from the same population, and $\bar{t}_1$ is the average coalescence time of two genes drawn from two different populations. Comparing Equations 5a and 10 reveals that σ_w^2 is proportional to $\bar{t}_0$ and σ^2 to $\bar{t}_1$. Consequently, the tree shown in Figure 5b not only minimizes SSD(WP) but also $\bar{t}_0$, the mean coalescent time within each population. Similarly, the MST shown in Figure 5c minimizes $\bar{t}_1$, the mean coalescent time of genes taken from different populations.

The determination of geographical clustering of gene trees, denoted as "intraspecific phylogeography" by AVISE *et al.* (1987), is of utility in describing organimal histories (*e.g.*, AVISE 1989; QUATTRO *et al.* 1991; VIGILANT *et al.* 1991), but the interpretation of population affinities from a single gene tree, even a unique best tree, is the subject of some ongoing discussion (LANGANEY *et al.* 1992; MADDISON, RUVOLO and SWOFFORD 1992; PAMILO and NEI 1988). The tree minimizing σ_w^2 will be the tree having the highest geographical consistency, as haplotypes found within the same population will tend to be as close to each other as possible. A method allowing computation of migration rates from phylogenetic information has recently been proposed (HUDSON, SLATKIN and MADDISON 1992; SLATKIN and MADDISON 1989), a method that uses the possible geographic inconsistencies of

any particular tree that is viewed as phylogenetic truth. Our procedure can be used to obtain an estimate of the maximum value of a geographic clustering index or the minimum number of migration events required to explain the current distribution of haplotypes among populations, and obtain their null distributions over the choice of possible trees. This would be quite useful in the study of gene flow patterns.

The authors thank ANDRÉ LANGANEY for his comments on the manuscript and two anonymous reviewers for their constructive criticism. L.E. was funded by FNRS Switzerland 32-28784.90 and 32-27845.89, and INSERM France 900 814, P.E.S. by NJAES/USDA-32102.

LITERATURE CITED

AVISE, J. C., 1989 Gene trees and organismal histories: a phylogenetic approach to population biology. Evolution **43:** 1192–1208.

AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, J. E. NEIGEL, C. A. REEB, and N. C. SAUNDERS, 1987 Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. **18:** 489–522.

BULMER, M., 1991 Use of the method of generalized least-squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8:** 868–883.

CANN, R. L., M. STONEKING and A. C. WILSON, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS, 1967 Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. **19:** 233–257.

CAYLEY, A., 1857 On the theory of analytical forms called tres. Phil. Mag. **13:** 172–176.

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

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.

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

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.

FELSENSTEIN, J., 1983 Parsimony in systematics: biological and statistical issues. Annu. Rev. Ecol. Syst. **14:** 313–333.

FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inferences and reliability. Annu. Rev. Genet. **22:** 521–565.

FELSENSTEIN, J., 1992 Phylogenies from restriction sites: a maximum-likelihood approach. Evolution **46:** 159–173.

FITCH, W. M., and E. MARGOLIASH, 1967 Construction of phylogenetic trees. Science **155:** 279–284.

GIBBONS, A., 1985 *Algorithmic Graph Theory.* Cambridge University Press, Cambridge.

HEDGES, S. B., S. KUMAR, K. TAMURA and M. STONEKING, 1992 Human origins and analysis of mitochondrial DNA sequences. Nature **255:** 737–739.

HUDSON, R. R., M. SLATKIN and W. P. MADDISON, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

KRUSKAL, J. B., 1956 On the shortest spanning subtree of a graph and the travelling salesman problem. Proc. Am. Math. Soc. **7:** 48–50.

LAKE, J. A., 1987 Rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. Mol. Biol. Evol. **4:** 167–191.

LANGANEY, A., D. ROESSLI, H. HUBERT VAN BLYENBURGH and P. DARD, 1992 Do most human populations descend from phylogenetic trees? Hum. Evol. **7:** 47–61.

LI, W.-H., 1986 Evolutionary change of restriction cleaving sites and phylogenetic inference. Genetics **113:** 187–213.

LI, W.-H., and M. GOUY, 1990 Statistical tests of molecular phylogenies. Methods Enzymol. **183:** 645–659.

LONG, J. C., 1986 Allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's *F* statistics. Genetics **112:** 629–647.

MADDISON, D. R., 1991 African origin of human mitochondrial DNA reexamined. Syst. Zool. **40:** 355–362.

MADDISON, D. R., M. RUVOLO and D. L. SWOFFORD, 1992 Geographic origins of human mitochondrial DNA: Phylogenetic evidence from control region sequences. Syst. Biol. **41:** 111–124.

MADDISON, D. R., and M. SLATKIN, 1991 Null models for the number of evolutionary steps in a character on a phylogenetic tree. Evolution. **45:** 1184–1197.

MERRIWETHER, D. A., A. G. CLARK, S. W. BALLINGER, T. G. SCHURR, H. SOODYALL, T. JENKINS, S. T. SHERRY and D. G. WALLACE, 1991 The structure of human mitochondrial DNA variation. J. Mol. Evol. **33:** 543–555.

NEI, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

NEI, M., and F. TAJIMA, 1983 Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105:** 207–217.

PAMILO, P., and M. NEI, 1988 Relationship between gene trees and species trees. Mol. Biol. Evol. **5:** 568–583.

PRIM, R. C., 1957 Shortest connection networks and some generalizations. Bell Syst. Technol. J. **36:** 1389–1401.

QUATTRO, J. M., J. C. AVISE and R. C. VRIJENHOEK, 1991 Molecular evidence for multiple origins of hybridogenetic fish clones (Poeciliidae:Poeciliopsis). Genetics **127:** 391–398.

QUATTRO, J. M., J. C. AVISE and R. C. VRIJENHOEK, 1992 Mode of origin and sources of genotypic diversity in triploid gynogenetic fish clones (Poeciliopsis:Poeciliidae). Genetics **130:** 621–628.

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

SANKOFF, D., 1983 Minimal mutation trees of sequences. SIAM J. Appl. Math. **28:** 35–42.

SLATKIN, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47:** 264–279.

SLATKIN, M., and W. P. MADDISON, 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

SMOUSE, P. E., and W.-H. LI, 1987 Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. Evolution **41:** 1162–1176.

SMOUSE, P. E., and J. C. LONG, 1988 A comparative F-statistics analysis of the Yanomama of lowland South America and the Gainj and Kalam of highland New Guinea, pp. 32–46 in *II International Conference Quantitative Genetics*, edited by B. S. Weir, G. Eisen, M. M. Goodman, and G. Namkoong. Sinauer Associates, Sunderland, Mass.

SMOUSE, P. E., J. C. LONG and R. R. SOKAL, 1986 Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool. **35:** 627–632.

SMOUSE, P. E., T. E. DOWLING, J. A. TWOREK, W. R. HOEH and W. M. BROWN, 1991 Effects of intraspecific variation on phylogenetic inference: a likelihood analysis of mtDNA restriction site data in cyprinid fishes. Syst. Zool. **40:** 393–409.

SOKAL, R. R., and F. J. ROHLF, 1962 The comparison of dendograms by objective methods. Taxon **11:** 33–40.

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., 1983 Convergent evolution and non-parametric inferences from restriction data and DNA sequences, pp. 151–179, in *Statistical Analysis of DNA Sequence Data*, edited by B. S. Weir. Marcel Dekker, New York.

TEMPLETON, A. R., 1992 Human origins and Analysis of mitochondrial DNA sequences. Nature **255:** 737.

THOMSON, E. A., J. V. NEEL, P. E. SMOUSE and R. BARRANTES, 1992 Microevolution in lower Central America: rare genes in an Amerindian complex. Am. J. Hum. Genet. **51:** 609–626.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES and A. C. WILSON, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.

WATTERSON, G. A., and H. A. GUESS, 1977 Is the most frequent allele the oldest? Theor. Popul. Biol. **11:** 141–160.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **1:** 323–334.

WRIGHT, S., 1965 The interpretation of population structure by F-statistics with special regards to systems of mating. Evolution **19:** 395–420.

Communicating editor: W. J. EWENS

APPENDIX

The restriction pattern and sample frequencies of 56 human mitochondrial DNA haplotypes are given in Tables 5 and 6 (pp. 358 and 359), respectively.

TABLE 5

Restriction pattern of 56 human mitochondrial DNA haplotypes

No.	Haplotype designation	Restriction pattern[a]
1	1	1111011111100011010111110111101000101010100110011101101100011 0
2	2	1111011111110011010111110111101000101010100110011101101100010 0
3	6	1111011111100011010111110111101000001010100110011101101100011 0
4	7	1111011111110011010111110111101000101010100110011101101100011 0
5	8	1111011111100011010111110111101000101010100010011101101100011 0
6	9	1111011111100011010111110111101000001010100010011101101100011 0
7	10	1111011111110011010111110111101100101010100110011101101100011 0
8	11	1111011111100001010111110111101100101010100110011101101100011 1
9	12	1111011111100011010111110111101000101011100110011101101100011 0
10	13	1111011111100011110111110111101000101010100110011101101100011 0
11	17	1111011111100011010111110111101000101010100110011001101100011 0
12	18	1111011111100011010111110111101000101010100110011111101100001 0
13	21	1111011111100011010111110111101100101010100110011101101100011 0
14	22	1111011111100011010111110111101100101010100110011101101100010 0
15	23	1111011111100011010111110111101000101010100110011101101100011 1
16	27	1111011111100011010111110111101000101110100110011101101100111 0
17	28	1111011111100011010111110111101000101010100110011101101100001 0
18	29	1111111111100011010111110111101000001010100010011101101100111 0
19	31	1111011111110011010111110111101100101010100110011101101100010 0
20	34	1111011111110011010111110111101000001010100110011101101100010 0
21	36	1111011111100011010111110111101000101010100100011001101100011 0
22	37	1111011111100011010111110111101100101010101110011101101100010 0
23	38	1111011111100111010111110111101000101010100110011101101100011 0
24	39	1111011111100011010111110111101000101110100110011101101100011 0
25	40	1111011111100011010111110111100000101010100110011101101100011 0
26	41	1111011111100011010111110111101000101010100110011101101101011 0
27	42	1111011111100011010111110111101000101010100111011101101100011 0
28	43	1111011111110011010111110111101000101010100100011101101100011 0
29	44	1111011111100011010111110111101000101010100110011101101100000 0
30	45	1111011111100011010111110111101000111010100110011101101100011 0
31	46	1111011111100001010111110111101000101010100110011101101100011 0
32	47	1111011111100011010111110111101000101010100110011101101100010 0
33	48	1111011111100011010111110111101000101010100110011101111100011 0
34	49	1111011111100011010111110111111000101010100110011101101100011 0
35	50	1111011111100011010111110111101000101010100111011101101100011 0
36	51	1111011111100011010111110111101000101010000110011101101100001 0
37	52	1111011111100011011111110111101000101010100110011101101100011 0
38	53	1111011111100011011111110111101000101010100110011101101100001 0
39	54	1111011111100011010111110111101000101010100110011101111100011 0
40	56	1111011111100011010111110111101000101010100110011101101100111 0
41	57	1111011111100011010111110111101000101010100100011111101100001 0
42	64	1111011111100011010111110111101000001110100110011101101100011 0
43	65	1111011111110011010111110111101000101010100010011101101100010 0
44	66	1111011111100011010111110111101000101010100110011111101100011 0
45	67	1111011111100011010111110111101000101110100110011101101100010 0
46	68	1111011111100011010111110111101001101010100110011101101100011 0
47	69	1111011111110011010111110111101000101010110110011101101100010 0
48	71	1111011111110011010111110111101010101010100110011101101100011 0
49	72	1111011111100011010111110111101000101010100100011101101100011 0
50	73	1111011111100011010111110111101000001010100110011111101100001 0
51	75	1111011111100011010111110111101000101010100110011001101100001 0
52	76	1111011111101011010111110111101000101010100110011101101100011 0
53	77	1111011111100011010111111111101000101010100110011101101100011 0
54	82	1111011111100001010111110111101100001010100110011101101100011 1
55	83	1111011111100011010111110111101000101010100100011001101100001 0
56	95	1101011111100011010111110111101000101010100110011101101100011 0

[a] Presence of a restriction site is coded by a zero and its absence by a one.

TABLE 6

Sample frequencies of 56 human mitochondrial DNA haplotypes in 10 populations

	Populations samples[a]									
Haplotypes	Tharu $N = 91$	Oriental $N = 46$	Wolof $N = 110$	Peul $N = 47$	Pima $N = 63$	Maya $N = 37$	Finnish $N = 110$	Sicillian $N = 90$	Israel Jews $N = 39$	Israel Arabs $N = 39$
1	48	32	23	11	59	30	87	50	15	22
2			39	19				3		1
6		1			2		2	9	14	1
7			29	12						1
8	2	2		2						
9	5	4								
10			2							
11							4		1	
12		2								
13	23	2								
17									1	
18							3	11		
21							8	1		
22									4	6
23								1		
27		1	2							
28	2	1								
29		1								
31										2
34				1				1		
36									1	
37									1	
38							2		1	
39			5	1	2				1	
40										1
41										1
42								1		1
43										1
44										1
45										1
46					2					
47	2					4	2	5		
48	2									
49	1									
50	1									
51	1									
52	2		2							
53	1									
54	1									
56								1		
57								2		
64			2							
65			1							
66			1							
67			1							
68			2							
69				1						
71			1							
72								1		
73								1		
75								1		
76								1		
77								1		
82							1			
83							1			
95						3				

[a] The original references of the population samples may be found in EXCOFFIER, SMOUSE and QUATTRO (1992).