Phylogenetic Analysis: Concepts and Methods

KENNETH K. KIDD[1, 2] AND LAURA A. SGARAMELLA-ZONTA[1]

INTRODUCTION

In a series of publications over the past several years, Cavalli-Sforza and Edwards have presented theoretical models and estimation procedures for the analysis of genetic relationships of human populations [1, 2, 3]. Their techniques of phylogenetic analysis have given a new perspective to the study of human variation and its origins. Moreover, their techniques are suitable for the study of microevolutionary relationships among populations of any species for which gene-frequency data are available. However, Cavalli-Sforza and Edwards only outlined the actual methodology involved. Recent studies concerning South American Indian tribes [4, 5] and domestic cattle breeds [6] have utilized their theoretical framework and their techniques; none of these papers discusses the methods thoroughly.

Two of the models have been compared in separate studies of human populations [4, 7]; although they used different methodologies, both concluded that the two models led to similar results. One study [7] was based on examination of all possible relationship structures, but involved only the one set of data. The other study [4] based the conclusion on examination of only an undisclosed fraction of the possible relationships. Thus, to date, there has been no systematic comparison of the various models proposed by Cavalli-Sforza and Edwards [3].

Because suitable gene-frequency data are rapidly accumulating and many workers are interested in such analyses, detailed discussion and comparison of the models and methods are in order. Here we present (1) the way in which Cavalli-Sforza and Edwards's theoretical models and estimation procedures can be applied to the study of real populations, (2) comparisons of the results obtained with the various methods used for the various models, and (3) our interpretation of the results of such an analysis. We also review some basic concepts.

GENETIC DISTANCES

The initial phase of the analysis is the transformation of population gene frequencies into genetic distances. To date, three slightly different distance measures

Received June 26, 1970; revised November 16, 1970.

This work was supported in part by a NSF graduate fellowship and a NIH postdoctoral fellowship to author Kidd and by grant GM 15422-03 from the National Institutes of Health. Use of the University of Wisconsin Computing Center was made possible through support from the National Science Foundation and the Wisconsin Alumni Research Foundation. Paper no. 1401 from the Laboratory of Genetics, University of Wisconsin.

[1] Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706.

[2] Present address: Department of Genetics, Stanford University School of Medicine, Palo Alto, California 94305.

have been used. All three start with the same single-locus transformation [8, 9]. The absolute angular distance θ_{ij} (measured in radians) at one locus between populations i and j is defined by

$$\cos \theta_{ij} = \sum_{k=1}^{m} \sqrt{p_{ik} \cdot p_{jk}}\,, \tag{1}$$

where m is the number of alleles at that particular locus, and p_{ik} and p_{jk} are the frequencies of the kth allele in populations i and j, respectively. Each pairwise distance between two populations is then calculated for the nth locus as

$$d_{ijn} = \frac{2}{\pi} \operatorname{arc\,cos} \sum_{k=1}^{m} \sqrt{p_{ik} \cdot p_{jk}}\,, \tag{2}$$

where the summation and symbolism are as in equation (1). Because of the multiplication of the arc cos by $2/\pi$, d_{ijn} is in gene substitution units with values ranging from 0 to 1. Additional explanations of the derivation have been published [6].

Cavalli-Sforza and Edwards proposed that in combining the pairwise distances obtained for several loci, the following formula be used for each pair of populations i and j:

$$G_{ij} = \left[\sum_{n} (\text{chord}^2_{ijn}) \right]^{1/2}, \tag{3}$$

where summation is over the number of loci and chord_{ijn} is an approximation of d_{ijn} at the nth locus given by

$$\text{chord}_{ijn} = \frac{2\sqrt{2}}{\pi} \left[1 - \sum_{k=1}^{m} \sqrt{p_{ik} \cdot p_{jk}} \right]^{1/2}, \tag{4}$$

where summation is the same as in equation (1). We refer to this as the geometric transformation [6]; the same transformation has been called "E" [10] and "root model" [4]. "Chord" was introduced [3] as an approximation to d_{ijn} in a Euclidean space. For small values of d_{ijn}, the chord approximation is very good and allows the construction of a space with the same scale, one unit per gene substitution. However, chord is a bad approximation for large values of d_{ijn}, since it ranges from 0 to only .9003 ($2\sqrt{2}/\pi$), not from 0 to 1 as stated by Fitch and Neel [4].

The necessity of maintaining a Euclidean space has been questioned and alternative transformations for combining the distance values obtained at single loci have been proposed. One additive transformation [2, 4] is

$$C_{ij} = \sum_{n} \text{chord}_{ijn}\,, \tag{5}$$

where summation is over the loci and chord_{ijn} is the value at the nth locus obtained from equation (4). An alternative additive transformation [6], not based on the chord approximations, is

$$A_{ij} = \sum_{n} d_{ijn}\,, \tag{6}$$

where summation is again over the loci and d_{ijn} is the value at the nth locus obtained from equation (2).

The distance measures which have been used in extensive analyses are then:

$$A_{ij} = \sum_n d_{ijn}\,, \qquad C_{ij} = \sum_n \text{chord}_{ijn}\,,$$

and

$$G_{ij} = \Big(\sum_n \text{chord}^2_{ijn}\Big)^{1/2}\,.$$

When these are applied to population gene frequencies, each of them yields a matrix of pairwise distances among the populations. **A, C,** and **G** are all symmetric matrices with zeros on the main diagonal. Because of the symmetry, usually only a triangular matrix of the potentially nonzero elements is considered.

These final values are not standardized with regard to the number of loci or with regard to the uniformity of allele testing in a given system. Therefore a locus must be represented in the data for all populations, and all alleles must be tested for in a consistent manner.

Other distance measurements are possible. Several of these have been critically discussed [10–14]. We wish to emphasize the use of distance matrices, rather than the various genetic models leading to the various measures or their statistical properties. Therefore, we have presented only the three transformations so far used on real population data and we are using only two of them in our analyses. We feel, however, that two of the other measures [11, 13] have strong genetic arguments in their favor and may well be preferable for future work.

Despite the mathematical differences between the A and G distances (particularly the approximation introduced by chord), it is worth noting that the results obtained separately from the two matrices are only slightly different [6, 15]. Also noteworthy is that the matrices of additive distances (**A** and **C**) are not necessarily representable in a Euclidean space. The minimum path method of analysis [3] requires the distances to be represented in a Euclidean space, and therefore **A** and **C** cannot always be used in that analysis, as we shall explain later.

TREES

The final results of our analyses are tree structures. We use trees, not only because of the precedent of evolutionary trees in biology, but also because we find them useful as relation graphs for depicting a set of complex interrelationships. The topology and the lengths of the various parts of the tree give, respectively, the qualitative and the quantitative relationships. Moreover, when dealing with populations represented in a multidimensional Euclidean space, we can draw the tree that connects them in that space on a two-dimensional surface by changing only the angles, while the topology and magnitude of the lines remain unchanged.

We follow the terminology of graph theory [16, 17] in order to define a tree: (1) a tree is a connected graph that has no circuits; (2) the degree of a point is the number of lines incident to it—a bifurcating tree has only points of degree one, $d(p_i) = 1$, which we call terminals, and points of degree three, $d(p_i) = 3$, which we call nodes; and (3) an unrooted tree has no point distinguished as different by being described as an origin or root. In our analyses we use only unrooted bifurcating trees, as specified by the above definitions. By limiting our analyses to bifurcating structures, we reduce the number of possible trees and greatly simplify the mathematics involved. Most of our methods give no information on a root or origin, thereby limiting us to unrooted trees.

When we superimpose a phylogenetic purport on this kind of relation graph, the definitions assume a precise biological meaning: (1) we do not allow hybridization between populations already separated; (2) each population splits only into two subpopulations; and (3) our analyses give no specific information about the position of the common ancestor, even if we can infer the probable root by other means. These trees are undirected graphs so long as no root exists. As soon as a segment of a tree is designated as containing the root, the tree becomes directed. Only in a directed graph can we say which populations are ancestral to which (i.e., specify the direction of evolution). A simulation study designed to test the relative merits of various methods [18] has shown that two types of cluster analyses can be used to assign a root to a tree with a higher probability of being correct than by chance alone. For a slightly different formulation of the problem [19], a very good maximum likelihood solution for the position of the root has been obtained. However, for the present we are examining only unrooted trees.

We have just defined two classes of points, the terminals and the nodes. Defining "segment" as the line connecting any two adjacent points, we can distinguish between terminal segments—those connecting a terminal to a node—and internal segments—those connecting two nodes. The terms "branch" and "arm" of other authors are synonymous with our "segment"; we use "branch" for a collection of connected segments.

From the preceding definitions, it follows that for N populations each tree has $2N - 2$ points (N terminals and $N - 2$ nodes) and $2N - 3$ segments (N terminal segments and $N - 3$ internal segments). There are

$$\prod_{K=3}^{N}(2K - 5)$$

different trees with N terminals for $N \geq 3$. Trees are different only if they cannot be superimposed with both topology and labeled terminals coinciding. Thus, for a single topological structure, several different trees are produced by permutations of the populations on the terminals. However, not all permutations produce different trees; whenever two topologically identical branches are joined to the same node, the number of different trees for that topology is reduced, as illustrated by the three trees in figure 1. Trees 1 and 3 are identical, whereas tree 2 is different.

MODELS AND ANALYSES

As the bases for the estimation of evolutionary trees, Cavalli-Sforza and Edwards proposed two models of evolution which we call the additive and the spatial models.

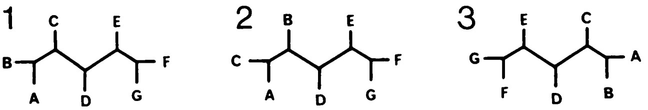

FIG. 1

The additive model states that the amount of evolution observed to separate two populations is equal to the sum of the amounts of evolution from those two populations to their common ancestral population. The spatial model states that evolution operates as a random walk and that with the appropriate measure of distances the populations can be placed in a Euclidean space in which the unit vector is a unit of evolution. Following these models, we obtain quantitative "solutions" for given tree structures and then choose that structure that allows the "best solution" according to some statistic calculated on the basis of the solution. The additive model has no rigorous statistical foundation for such solutions, while the spatial model has a rigorous statistical basis with evolution considered as a Brownian motion/Yule process. Unfortunately, the use of maximum likelihood to estimate the relevant parameters of this process has proved intractable [20]. Thus there is little or no statistical theory to justify the "solutions" obtained following either model.

The adoption of either model implies that evolution occurred independently in each ancestral population. The measures of genetic distance seem less valid in trying to reconstruct ancestry by these models than in simply measuring differences among extant populations. Indeed, in the case where the additive model might seem most likely to hold—namely, the evolutionary change of the amino-acid sequence of a protein—the observed distances between the species do not allow an exact reconstruction by the additive model because examples of parallel, reverse, and convergent evolution are observed [21, 22].

In spite of these problems, analyses assuming the additive model have been shown to produce results agreeing with available historical data [6]. The method of minimum path (described later) gives an intuitively-based solution to the spatial model and has similarly been shown to give reasonable results [5, 15]. We are therefore accepting the methods of analysis for both the additive and spatial models and will explain their mathematical bases and practical applications in more detail.

Additive Model

Cavalli-Sforza and Edwards [1] proposed the least-squares method to solve the additive model; Fitch and Neel [4] used a computer program based on weighted averaging. Because of its simple formulation in matrix notation and its statistical properties, we prefer least squares and will discuss only that method.

For N populations, there are $N(N-1)/2$ pairwise distances and $2N-3$ segments (see preceding section on trees). We can write a system of equations, in matrix notation,

$$\mathbf{D} = \mathbf{BS}, \tag{7}$$

where **D** and **S** are, respectively, the distance vector $\{[N(N-1)/2] \times 1\}$ and the segment vector $[(2N-3) \times 1]$, and **B** is a [0, 1] matrix $\{[N(N-1)/2] \times (2N-3)\}$ representing the form of the tree by specifying for all pairs of populations (rows) the presence (1) or absence (0) of each segment (column) in the pathway connecting them; there is a unique **B** matrix for every tree. The data consist of the values in the **D** vector, the **B** matrix is dictated by the tree structure we wish to solve, and the **S** (segment) vector is unknown and to be estimated.

The system of equations, (7), is solved for **S** by

$$\mathbf{S} = (\mathbf{B}'\mathbf{B})^{-1}\,\mathbf{B}'\mathbf{D}\,, \tag{8}$$

as given in most linear algebra texts. An exact solution is unlikely with real data and this estimation of **S** has the desirable quality that

$$(\mathbf{B}\hat{\mathbf{S}} - \mathbf{D})'(\mathbf{B}\hat{\mathbf{S}} - \mathbf{D}) = \Sigma(\text{error}^2) \tag{9}$$

is minimized. Following Cavalli-Sforza and Edwards, we can use the quantity $\Sigma(\text{error}^2)$ as a measure of the deviation of the system of equations from exact additivity.

The least-squares solution can be applied only to unrooted trees because $(\mathbf{B}'\mathbf{B})$ must be nonsingular and this only holds for unrooted trees. If a root is inserted, $(\mathbf{B}'\mathbf{B})$ always has two equal rows (and columns)—those corresponding to the two segments incident at the root [d(root) = 2]. The rank of $(\mathbf{B}'\mathbf{B})$ is $2N - 3$ = number of unknowns = number of segments in an unrooted tree.

After obtaining quantitative solutions for different trees, the discrimination among them is based primarily on their relative ability to yield an additive solution. The tree allowing the most nearly additive solution is considered best. Only two of several possible ways to measure deviation from precise additivity have so far been used: the $\Sigma(\text{error}^2)$ value of least squares [2] and the "%SD" value [21]. In addition to deviation from additivity, Cavalli-Sforza and Edwards impose a further restriction by considering as unacceptable any tree whose solution has negative segments. Evidence [23] linking the deviation from additivity, $\Sigma(\text{error}^2)$, with lack of negative segments has led to a new statistic, related to both, that will be presented here.

Although Cavalli-Sforza and Edwards never published reasons for their rejection of trees with negative segments, such rejection has been their policy [2]. In fact, it seems possible that the imposition of an incorrect topology on an essentially additive set of data results in a solution containing one or more negative segments even when the original (or correct) tree has only positive values. This view is supported by an examination of exactly additive data. We examined all 105 possible trees for six populations using two different sets of data. In each case, only the original tree used to generate the distance matrix allowed an all-positive solution, the exactly correct solution. Real population data do not allow an exact solution because of statistical fluctuations. However, the present results—see discussion of least-squares method (LS) length and of the changing subroutine—also support empirically the rejection of trees containing negative segments, even for real population data. Thus, we reject trees containing negative segments.

An alternative approach to discrimination among trees is used for the minimum-path method (explained below) and is also applicable to the additive solutions. It considers that the total amount of evolution given in the relationships should be a minimum. By this we do not imply that evolution proceeds in a directed way. Quite the contrary, our total approach assumes that evolution proceeds in a random manner with respect to the microevolutionary differences we are observing. Even though a complete demonstration is lacking, it does seem reasonable that under these conditions the best estimate of the actual evolutionary relationships is the one which requires the least evolution. When we use the additive model, we will base our estimate of parsimony on the results of the least-squares solution.

For each different tree solved by least squares, $\hat{\mathbf{S}}$ (the vector of segment lengths) is estimated in such a way that the $\Sigma(\text{error}^2)$, as in equation (9), is a minimum. As a measure of parsimony of evolution, we define a new statistic based upon $\hat{\mathbf{S}}$:

$$L(\hat{\mathbf{S}}) = \sum_{i=1}^{2N-3} |\hat{s}_i| \;, \tag{10}$$

where $\hat{s}_i$ are the elements of $\hat{\mathbf{S}}$. The absolute value is used, since even negative values are considered as amounts of evolution. The statistic $L(\hat{\mathbf{S}})$, called LS length, does not produce orderings of trees the same as $\Sigma(\text{error}^2)$. We are unable to give mathematical explanations for several interesting aspects of this statistic. We are therefore presenting conjectures followed by some of the results upon which they are based.

CONJECTURE 1. If there exists at least one tree whose least-squares solution has no negative segments, the tree with the minimum LS length has no negative segments.

CONJECTURE 2. The tree with the lowest $\Sigma(\text{error}^2)$ value among those with no negative segments is the tree with the minimum length.

In analyses of many sets of data, we have always found both conjectures to hold if all trees were examined. When only a fraction of the total number of trees were examined [15], we have always observed the first conjecture to hold, but have not always found the second conjecture to hold among those trees examined. The second conjecture, if valid, could be valuable as an aid in evaluating results on a small sample of the possible trees—if no all-positive tree shows both the minimum $\Sigma(\text{error}^2)$ and minimum LS length, the "best" tree has not been found. However, the converse is *not* implied by the conjecture.

For comparison with the other statistics just discussed, we also calculated a "%SD" [4, 21] on the trees solved by least squares. However, we normalized to the expected pairwise distances, rather than the observed, because least-squares estimates were used. No algorism is known for finding the tree with the absolute minimum value for any of these statistics save evaluation of all possible trees.

Spatial Model

The intuitive solution for the spatial model [24] is here referred to as the minimum-path method. Edwards and Cavalli-Sforza presented it as the minimum evolution method, implying that the best estimate of evolution might be the path invoking the minimum amount of evolution. It is considered an approximation to the maximum-likelihood solution for a Brownian motion/Yule process, but is independent of the assumption of a Yule process. The populations are placed in a Euclidean space and the pathway sought is the shortest net connecting the populations in that space. In contrast to the least-squares method, here the process of solving any given structure is iterative, and by successive approximations it finds the minimum length possible for that particular tree.

The terminals are plotted in a Euclidean space at positions determined by the data and are then fixed at these positions. The nodes are initially placed at the origin. At each iteration, the coordinates of each node in turn are estimated by considering that node as the "Steiner point" for its three adjacent points [3]. This point is calculated as a center of gravity with appropriate weights assigned to the three points. The derivation of the appropriate weights and the original program from which ours is derived are by Edwards (unpublished). Iteration proceeds until the desired degree of convergence is obtained. Note that the minimum-path solution can be applied only to unrooted trees: for each node three adjacent points are necessary to calculate its coordinates; a root has only two adjacent points in a bifurcating tree and, therefore, disappears into the straight line connecting the two points on either side.

The minimum length required for a particular network (here called minimum path method [MP] length to distinguish it from our similar LS length) is the measure of the goodness of that structure and is used to discriminate among the various trees. It corresponds to

$$\text{MP length} = \sum_{i=1}^{2N-3} s_i \,, \tag{11}$$

where s_i is the length assigned to the ith segment. Since these segment lengths are in a Euclidean space, the minimum possible value of each is 0. Under these conditions, equations (10) and (11) are identical. The shortest network will be "stable" [3; A. W. F. Edwards, personal communication] (i.e., will have no segments of zero length except those generated by nodes coinciding with populations). Therefore, we consider trees with zero length internal segments unacceptable unless caused by two nodes coinciding with a population (a phenomenon we have not observed). Otherwise, the tree with the lowest MP length among those evaluated is the "best." As with the statistics for the additive analysis, no algorism is known for finding the tree with the absolute minimum MP length, save the evaluation of all possible trees.

The minimum-path method has one advantage over the least-squares method—it can be used to estimate the gene frequencies of the ancestral populations. For such estimations, the population coordinates are the square roots of the allelic frequencies, each allele allotted a Cartesian axis. This is the same space in which the geometric distances are calculated. However, we would be interested in such estimates only for the better trees obtained, and the calculation of the minimum-path solution is faster if the number of dimensions is reduced by a translation and rotation of the axes. This is possible if there are more alleles than populations—N populations are always representable in only $N - 1$ dimensions, if they are originally in a Euclidean space of higher dimension. In practice, an arbitrary coordinate system of $N - 1$ dimensions is calculated from the pairwise distance matrix and does not require knowledge of the original coordinates. Because the distance matrix is sufficient, it is possible to use any distance matrix that can fit into a Euclidean space. When distances are calculated in some fashion not involving a Euclidean space, a Euclidean space in which they fit has no clear relationship to the gene frequencies, nor is it an "evolutionary space" or "character space" as is obtained from the geometric transformation.

Fitch and Neel [4] state that only the geometric (their "root") distances may be used for minimum-path analyses. For their data this was true, but it is not generally true. The only mathematical restriction in the use of the minimum-path method is that the distance matrix be representable in a Euclidean space—a condition occasionally fulfilled by the **A** and **C** matrices. It is not generally apparent, a priori, whether a distance matrix will fit into a Euclidean space, since this is determined by inequalities. The triangle inequalities for the distances connecting any three points (each distance must be greater than or equal to the absolute difference of the other two and less than or equal to their sum) must hold, but they are not sufficient. An illustration requires at least four points and three dimensions. If the triangle inequalities hold for all the subsets of three, four points can always be placed in three dimensions using any five of the six pairwise distances. These then form two triangles with one side in common. The sixth distance, which must connect the two "opposite" vertices to give a general tetrahedral form, is then constrained to lie between, or to equal, a certain minimum value and a certain maximum value. The lower limit on the sixth value occurs when the two triangles lie in the same plane and are "superimposed"; the upper limit occurs when they are in the same plane but adjacent. If the sixth distance value is outside this range, no Euclidean representation is possible.

Comparisons of Models and Statistics

The two methods, least squares and minimum path, produce solutions that embody different mathematical properties and hence are not comparable in the segment lengths produced, or, necessarily, in the orderings of the trees. It has been shown [3] that the least-squares method yields segment lengths too short to connect the populations in the space in which the minimum-path tree lies. This difference is reflected in smaller values for the statistic LS length as opposed to the MP length. In fact, none of the statistics described here are numerically comparable. Comparisons are possible only among trees evaluated on the same distances by the same statistic. However, we have compared the rankings of trees by the various statistics in two ways—by correlations between them and by comparisons of the positions of the all-positive least-squares solutions in the distributions.

Using several sets of data (some completely independent and some partially independent), we have analyzed the distance matrices. In all cases, we have obtained similar methodological results. We shall discuss the results of four independent sets of data as representative. Data sets 1 and 2 are, respectively, a **G** and a **Ds** (based on anthropometric measurements) distance matrix. Elsewhere we present the distances and biological interpretations for these human populations on Bougainville Island [15]. Data set 3 is a **G**-distance matrix [5] based on 11 loci for seven South American Indian tribes. Data set 4 is an **A**-distance matrix for seven domestic cattle breeds [6, 25]. Figure 2 gives the sets of distributions of trees evaluated according to the methods and statistics just discussed for the first of these data sets. The distributions for the other three data sets have been deposited with a documentation service and are available on request.*

The distributions in figure 2 are complete in that all 945 trees for seven populations were solved by least squares and minimum path. The distributions are then based on the three separate statistics ($\Sigma[\text{error}^2]$, "%SD," and LS length) for the least-squares solution of the additive model and on the MP-length statistic for the minimum-path

* National Auxiliary Publications Service of American Society for Information Science, c/o CCM Information Corp., 909 Third Avenue, New York, New York 10022.

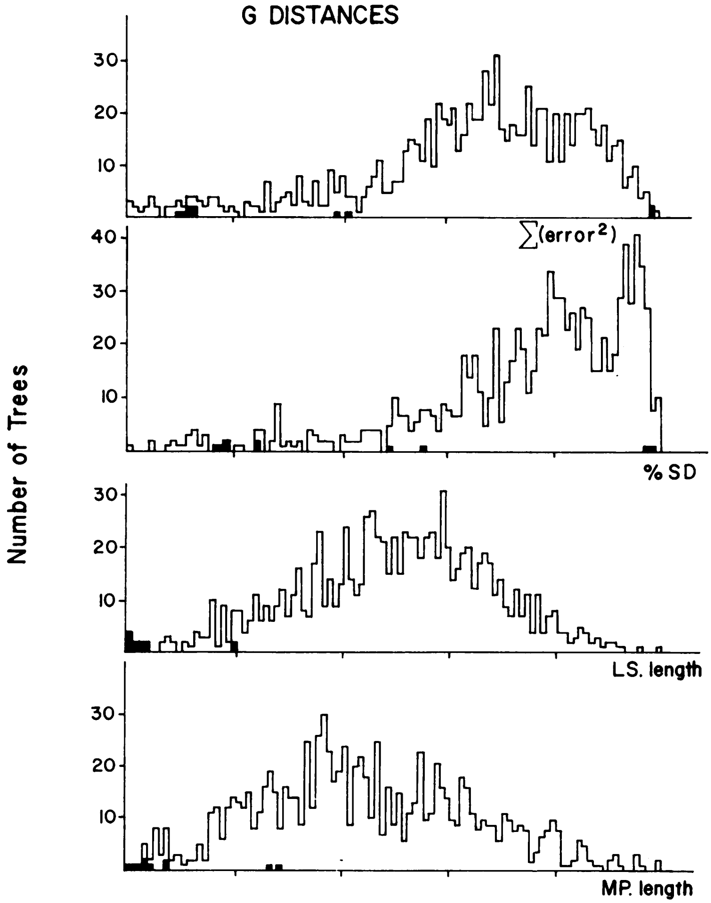

FIG. 2.—Distributions produced by different statistics on all 945 possible trees relating seven human populations on Bougainville Island. The data used and biological interpretations are presented elsewhere [15] as set iv. The first three distributions are produced by three different statistics applied to the least-squares solution, the fourth by the MP length statistic applied to the minimum-path solution (see text for description of methods and statistics). Since the values of the different statistics are not numerically comparable, the range of each is arbitrarily divided into 100 intervals along the abscissa; the number of trees in each interval is plotted on the ordinate. The darkened points indicate those trees that had no negative segments when solved by least squares. The number of negative segments is a property of the least-squares solution; the position of that tree is determined by the statistics. Thus the same trees are marked in all distributions, even in the MP length distribution which is not based on least squares. The respective ranges of the statistics are $\Sigma(\text{error}^2)$, 0.01375–0.07015; "%SD," 6.0773–15.013; LS length, 1.2701–1.8250; and MP length, 1.6522–2.1084.

solution. The darkened points in the histograms are the positions of the trees whose least-squares solution had no negative segments. To facilitate comparison, these same trees are also marked in the distribution according to MP length, even though it is not based on the least-squares solution. We have not marked the trees lacking zero-length segments in the minimum-path analysis since they are too numerous.

The most striking aspect of these distributions is the consistency of the form of each statistic's distribution over the sets of data; all show forms similar to those in figure 2. No formulas for the generation of these distributions exist, but these results and others [6, 7] invite an attempt to derive such formulas. It is clearly desirable to have the distribution functions, since the range of the statistic could then be estimated from the distance matrix or from a small random sample of trees. It is especially important to be able to compare a single tree with an expected minimum value when it is impossible to examine all trees, as we shall discuss shortly.

The next most striking aspect of these distributions is the relative positions of the all-positive trees. Our new statistic (LS length) yields a distribution with the all-positive trees concentrated into the lower part of the distribution. In figure 2, for example, two all-positive trees are among the three trees with highest Σ(error^2); however, when ranked by LS length, these same two trees are among the lowest 7%. In the distributions for the other data sets, the shift is equally obvious. We have observed this shift in all cases where we have examined all the possible trees. The tree with the lowest LS length has always been the all-positive tree with the lowest Σ(error^2).

Table 1 gives the correlations of these various statistics obtained from the four data sets. Each correlation is calculated using the values of the statistics for all 945 possible trees; however we cannot consider the trees as independent and hence cannot directly assign significance levels to these correlations. Each data set is independent, however, and it is possible to consider these correlation coefficients as independent estimates, calculate a mean and standard error, and base our conclusions on these. In practice, we have calculated the mean *z* value, its standard error, and the *t* value. With the exception of the correlation of "%SD" with MP length, all are significantly different from zero. The "%SD" statistic is highly correlated with Σ(error^2), much more than visual comparison of the two distributions indicates; LS length and MP

TABLE 1

CORRELATIONS OF STATISTICS USED TO RANK TREES

STATISTICS	DATA SETS				MEAN *z*	SE
	1	2	3	4		
Σ(error^2)-"%SD"	.962	....	.840	.947	1.665	.227*
Σ(error^2)-LS length	.195	.514	.250	.536	0.405	.104*
Σ(error^2)-no. negative segments	.287	.401	.329	.312	0.346	.028**
Σ(error^2)-MP length	.366	.577	.305	....	0.452	.105*
"%SD"-LS length	.304	.465	.297	.630	0.466	.102*
"%SD"-no. negative segments	.341	....	.329	.334	0.348	.004***
"%SD"-MP length	.444	....	.317	....	0.403	.074
LS length-no. negative segments	.662	.585	.560	.567	0.686	.038***
LS length-MP length	.822	.837	.799	....	1.157	.033***
No. negative segments-MP length	.595	.524	.450	....	0.584	.058**

NOTE.—The data sets and statistics are described in the text except for "no. negative segments" which is the total number of negative segments in the least-squares solution of a tree. Except for MP length, all statistics are based on the least-squares solution. The significance of the correlation coefficients is difficult to assess directly. However, since each data set provides a completely independent estimate of the correlation between two statistics, we have calculated the significance levels by means of a *t* test on the *z* values. Only the correlation between "%SD" and MP length is not significantly different from zero.

* $P < .05$. ** $P < .01$. *** $P < .001$.

length are also highly correlated. Although of moderate value and not significantly different from each other, the correlation of LS length is higher with the number of negative segments than with $\Sigma(\text{error}^2)$. The remaining correlations are low to moderate.

The high correlation of LS length and MP length is also shown graphically in figure 3, using a set of six populations [15] representable in a Euclidean space using both the **A** and **G** distance matrices. The similarity and the high linear correlation of the results obtained by the two methods are obvious.

In all of these sets of populations, the most biologically (and in the case of the cattle breeds, historically) meaningful trees have been those with lowest LS length. These closely correspond to the all-positive trees with low $\Sigma(\text{error}^2)$ and to the trees with lowest MP length. The biological interpretations are published elsewhere, and only the analyses of data set 3 require comment here. These analyses confirm the biological results of Ward and Neel [5] obtained from the minimum-path analysis of only some of the possible trees. Our complete analyses of their 5-loci and 6-loci distance matrices similarly confirm their results for those data sets. In all three cases, the trees they illustrated [5] were not the trees we obtained as best, but were only minimally different. We attribute this discrepancy partially to their not examining all trees, and partially to our using distances that had been rounded off to three figures. The differences are probably not statistically significant and are biologically unimportant.

Since it is not obvious from the results presented here, it is worth mentioning that

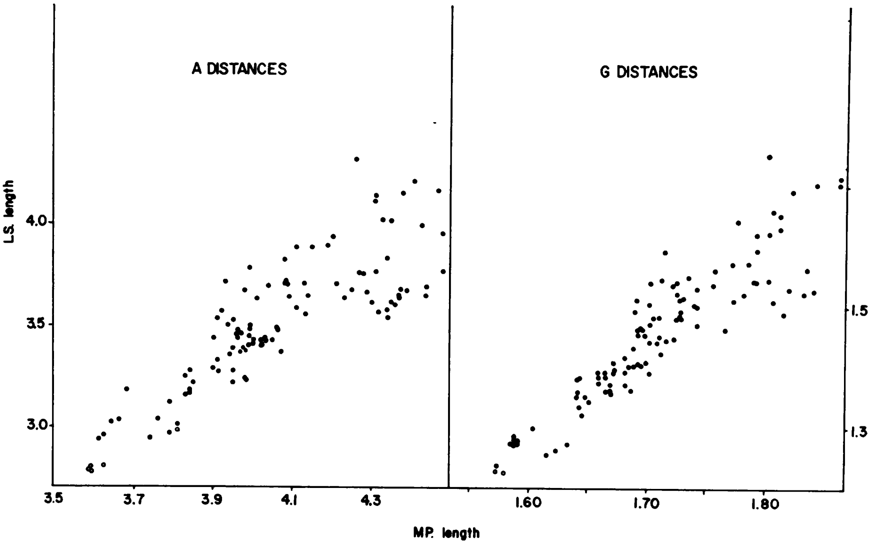

FIG. 3.—Scatter diagrams for all 105 trees for six human populations on Bougainville Island. The data used and the biological interpretation of these results are published [15] as set i. The *A* and *G* distances were analyzed by both least squares and minimum path. Each tree is plotted according to its value for the two statistics, LS length and MP length. Those trees that had all-positive least-squares solutions are plotted as open circles. These plots demonstrate the high linear correlation of the two methods using these two statistics.

the form of the better trees is dependent on the data. Different sets of data subjected to the same analyses yield different "best" trees. This is obvious from the examination of "good" trees from these distributions as shown in the papers cited.

Extension to Larger Numbers of Populations

The approach just described is limited to a small number of populations. With more than seven populations, it is currently too costly and time-consuming to evaluate all possible trees; it is only possible to evaluate a fraction of the total number of different trees. The problem becomes that of choosing the set of trees most likely to contain the best structures. With trees picked at random, the probability of finding the best possible trees is directly related to the number evaluated. As the number of populations increases, quite obviously the number of trees that can be evaluated in a given amount of time decreases. Picking trees at random therefore gives increasingly diminishing returns. Based on our analyses [6, 15], we have concluded that two methods of generating trees are very useful and largely free from subjective bias. Moreover, we feel they give a high probability of finding the best tree, although we cannot yet quantify this probability.

First, we have used the method of cluster analysis [26] which divides the populations into the two clusters with the minimum within-cluster variance. Repeated application of this method to each cluster produced by the previous cycle generates a bifurcating tree. We use the several best clusterings at each cycle to obtain a family of likely trees. These trees are then evaluated and compared. This method for cluster analysis is well covered in the literature [4, 6, 26, 27].

Second, we have used a technique [6] which slightly alters these trees to produce still more trees for evaluation and comparison. It is essentially the same method that Cavalli-Sforza and Edwards used in their work (see the program for the minimum-path analysis distributed by Edwards). It assumes that negative segments (zero-length segments in the minimum path) are indeed the result of choosing an improper topology and are the best places to alter the tree. In addition to these two methods, likely trial trees should result from many other methods used to generate trees [27–30], some of which can be adapted to produce families of trees rather than a single structure.

Because of its usefulness and the relationship it demonstrates between negative segments and deviation from additivity, we present here some of the results obtained with our program for successive changes of tree structure. Briefly, the program operates by searching the least-squares solution of a tree for negative segments. When a negative segment is found, the topology of the tree is changed around that segment. The new tree is then evaluated by least squares and the segment around which the change was made is checked. If the segment is positive, the program then searches the tree for negative segments, thus restarting the cycle with this new tree. If instead the segment is still negative, the remaining possible topological change is made around the segment, and the new tree evaluated by least squares. We have always found that one of the three possible arrangements allows a positive solution for the relevant segment. This procedure ends when no negative segments remain.

Figure 4 shows the most convincing illustration of the usefulness of this method.

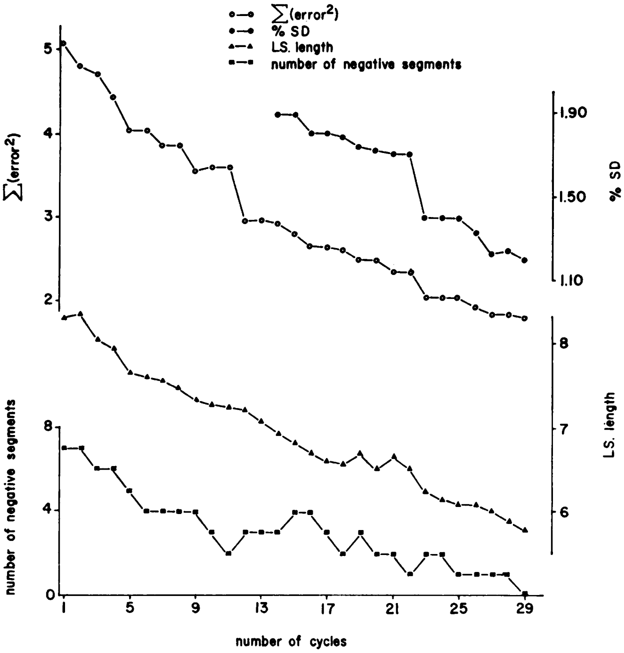

FIG. 4.—Improvement resulting from the application of the change subroutine. Each point on the abscissa represents a different tree, each produced by the elimination of a negative segment from the previous tree by changing the tree structure [6]. The statistical evaluations of each tree are given by the four graphs according to their individual scales. The input tree (cycle 1) was a random tree for 19 human populations [15].

The starting point for this series of successive alterations was a random tree relating 19 populations on Bougainville [15]. These statistics decline steadily over the 29 cycles; only the total number of negative segments shows an irregular fluctuation. The input tree was the worst or nearly worst tree found by all statistics. The all-positive tree that resulted from this series was by all statistics very close to the best obtained among the 270 trees examined; the relationships shown are also very similar to those in figure 2 of Friedlaender et al. [15].

We have used this subroutine on many sets of data and many different input trees, and have always obtained a similar improvement in the fit of the trees from successive cycles. Usually, however, we find an all-positive tree in only a few cycles, and the improvement is less dramatic than illustrated here. It is also possible, after once arriving at an all-positive tree, to continue altering around segments less than a small positive value. This occasionally gives other all-positive trees directly, but more commonly initially yields trees with negative segments, from which, in turn, other all-positive solutions are occasionally found. There is, however, very erratic behavior of the statistics, and occasionally closed loops of changes occur. A limited number of such changes are useful, but the method is not systematic and is stopped after an arbitrary small number of changes.

DISCUSSION AND CONCLUSIONS

The distance measurements used here (*A* and *G*) are very similar [6, 15] and have produced results that are largely equivalent. We expect that other distance measurements [11, 13] will show a similar high correlation with these distances. Therefore it should be sufficient to use only one measure of distance.

Our analyses show that the statistics $\Sigma(\text{error}^2)$ and "%SD" are highly correlated, as are LS length and MP length. Although most comparisons between statistics in those two groups are significantly positive, it is nonetheless evident from the much lower values that the pairs are quite different. It is interesting to note that the number of negative segments is significantly positively correlated with all four statistics. Of course, one cannot conclude from this which method should be used, but only one statistic from each pair should suffice.

In deciding which method or statistic to use, additional information is required. We prefer LS length and MP length because the results obtained have been in closer agreement with historical data. This preference is supported by the initial results of a simulation study [18] designed to study the errors involved in estimation of phylogenetic trees by these methods. Of the two similar methods (least-squares solution using LS length and minimum-path solution using MP length), we prefer the former for a very pragmatic reason: it has consistently taken less computer time. This appears to be the result of the very slow convergence, and hence large number of iterations, for some particular networks. While it may be possible to improve the minimum-path program, it remains the less efficient method for the present.

The method for changing topologies is very useful and demonstrates the likely validity of the original assumption that negative segments are not true representations of evolutionary relationships. It may be that it will not be possible to represent the correct relationships with an all-positive tree for some populations, but the basic assumption of random processes, if valid, makes this unlikely.

At present there is no way to estimate the variance of the different statistics used to rank trees. This is especially important as the original data are only estimates of the gene frequencies of the populations. Therefore, interpreting the results requires some caution. We reject the idea that a single best tree can adequately represent a distance matrix. When we evaluate all possible trees, we see no statistical basis for identifying a level of significance which would establish one or a small group of trees as being better than the rest. The problem is even more acute when we analyze only a fraction of all the possible structures, since we have no assurance in this case that the trees from the lower end of the distribution have been evaluated. This problem can be partly overcome by examining the several better trees obtained. We can then consider most definitive those relationships which are constant in all trees, and consider most indecisive and suspect those relationships which differ among the trees.

Because the distributions according to LS length are approximately "normal," it has been suggested (L. L. Cavalli-Sforza, personal communication) that a rough indication of the expected minimum value of the statistic LS length may be obtained by considering the distribution to be normal. A few random trees could be evaluated and used to estimate the mean and variance. The number of trees is known and hence the extreme values could be estimated. While this involves several approximations, it is worth future investigation, since we might then have some assurance that we were basing our conclusions on trees from the lower end of the total distribution, even though only a small fraction had been evaluated.

Cavalli-Sforza and Edwards originally presented precise models of evolution; on the basis of these they developed estimation procedures. Since estimation by maximum likelihood has been shown to be practically impossible, they found it necessary to fit these models to real data by approximate estimation methods, which are not completely satisfactory because they must rely on intuition to obtain solutions. A quantitative estimation of the errors involved in using these methods will have to be obtained from simulation studies, now in progress [18]. In practice, frequent deviations from the original models occur (e.g., our use of the minimum-path solution for additive distances). It is encouraging to note, therefore, that the various applications of these methods to population data [5, 6, 15] have given results that agree well with the known histories and relationships of the various groups. These studies provide a pragmatic basis for using these methods to estimate microevolutionary relationships and to quantify relative amounts of divergence among populations.

When additional information about the populations being studied confirms the likely validity of the assumptions underlying the models, the trees reconstructed by these methods are estimates of the phylogenetic relationships of the populations. In situations where the original assumptions do not hold, we cannot reconstruct and represent the actual evolutionary history of the populations with trees, but instead we can consider the trees as synthetic representations of genetic similarities.*

SUMMARY

The examination of the various methods and criteria for evaluating evolutionary trees based on genetic distances shows that all are positively correlated. Comparison of the results with the known relationships in the various sets of data indicates that two methods are more likely to give biologically correct results—the minimum-path method which uses a minimum evolution criterion, and the least-squares method when a minimum evolution statistic is also used. These results provide a pragmatic basis for the application of these methods to populations of unknown relationships to elucidate their evolutionary or genetic relationships.

ACKNOWLEDGMENTS

We wish to thank C. Matessi, S. D. Jayakar, J. F. Crow, A. W. F. Edwards, and L. L. Cavalli-Sforza for their advice and criticism. We also thank R. H. Ward and J. V. Neel for making their data available before publication. Particular thanks is due to J. S. Friedlaender for his many helpful comments during the course of the work and preparation of the manuscript.

* Descriptions of the computer programs used in this and other studies [6, 15, 25] are being prepared. The programs are written in FORTRAN V for the Univac 1108, but require few modifications for use with FORTRAN IV compilers. The programs will be available from either author or from J. S. Friedlaender, Anthropology Department, Harvard University.

REFERENCES

1. Cavalli-Sforza LL, Edwards AWF: Analysis of human evolution, in *Proceedings 11th International Congress of Genetics* 3:923–933, 1964
2. Cavalli-Sforza LL: Population structure and human evolution. *Proc Roy Soc* [Biol] 194:362–369, 1966
3. Cavalli-Sforza LL, Edwards AWF: Phylogenetic analysis: models and estimation procedures. *Amer J Hum Genet* 9:234–257, 1967; also in *Evolution* 21:550–570, 1967
4. Fitch WM, Neel JV: The phylogenic relationships of some Indian tribes of Central and South America. *Amer J Hum Genet* 21:384–397, 1969
5. Ward RH, Neel JV: Gene frequencies and microdifferentiation among the Makiritare Indians. IV. A comparison of a genetic network with ethnohistory and migration matrices; a new index of genetic migration. *Amer J Hum Genet* 22:538–561, 1970
6. Kidd KK: Phylogenetic analysis of cattle breeds. Ph.D. thesis, Univ. Wisconsin, 1969
7. Zonta LA: Analisi di alberi evolutivi in popolazioni umane. *Atti Ass Genet Ital* 11:326–335, 1966
8. Bhattacharyya A: On a measure of divergence between two multinomial populations. *Sankhyā* 7:401–406, 1946
9. Cavalli-Sforza LL, Conterio F: Analisi della fluttuazione di frequenze geniche nella popolazione della Val Parma. *Atti Ass Genet Ital* 5:333–343, 1960
10. Balakrishnan V, Sanghvi LD: Distance between populations on the basis of attribute data. *Biometrics* 24:859–865, 1968
11. Cavalli-Sforza LL: Human diversity, in *Proceedings 12th International Congress of Genetics* 3:405–416, 1969
12. Cavalli-Sforza LL, Zonta LA, Nuzzo F, et al: Studies on African Pygmies. I. A pilot investigation of Babinga Pygmies in the Central African Republic (with an analysis of genetic distances). *Amer J Hum Genet* 21:252–274, 1969
13. Edwards AWF: Distances between populations on the basis of gene frequencies. Submitted to *Biometrics,* 1970
14. Edwards AWF, Cavalli-Sforza LL: Affinity as revealed by differences in gene frequencies, in *The Assessment of Affinity between Human Populations,* edited by Wiener JS, Oxford, Oxford Univ. Press. In press, 1971
15. Friedlaender JS, Sgaramella-Zonta L, Kidd KK, et al: Biological divergences in south-central Bougainville: an analysis of blood polymorphism gene frequencies and anthropometric measurements utilizing tree models, and a comparison of these variables with linguistic, geographic, and migrational "distances." *Amer J Hum Genet* 23:253–270, 1971
16. Ore O: *Graphs and their Uses.* New York, Random House, 1963
17. Cotterman CW: Combinatorial problems in genetics. Unpublished notes, 1966
18. Kidd KK, Cavalli-Sforza LL: Number of characters examined and error in reconstruction of evolutionary trees, in *Proceedings of the Anglo-Romanian Conference on Mathematics in the Historical and Archeological Sciences,* edited by Kendall DG, Hodson FR, Tăutu P, Edinburgh, Edinburgh University Press. In press, 1971
19. Haigh J: The recovery of the root of a tree. *J Appl Probability* 7:79–88, 1970
20. Edwards AWF: Estimation of the branch points of a branching diffusion process. *J Roy Statis Soc B* 32:155–164, 1970
21. Fitch WM, Margoliash E: Construction of phylogenetic trees. *Science* 155:279–284, 1967
22. Fitch WM, Margoliash E: The construction of phylogenetic trees. II. How well do they reflect past history? *Brookhaven Sympos Biol* 21:217–242, 1968
23. Kidd KK, Sgaramella-Zonta LA: A test of an additive model for phylogenetic analysis (abstr.). Genetics 61:s31–s32, 1969

24. Edwards AWF, Cavalli-Sforza LL: The reconstruction of evolution (abstr.). *Ann Hum Genet* 27:105, 1963
25. Kidd KK, Sgaramella-Zonta LA: Relationships of domestic cattle breeds, in *Proceedings 12th International Conference on Animal Blood Groups and Biochemical Polymorphisms*, Budapest, Hungary. In press, 1971
26. Edwards AWF, Cavalli-Sforza LL: A method for cluster analysis. Biometrics 21:362–375, 1965
27. Gower JC: A comparison of some methods of cluster analysis. *Biometrics* 23:623–637, 1967
28. Boyce AJ: Mapping diversity: a comparative study of some numerical methods, in *Numerical Taxonomy*, edited by Cole AJ, New York, Academic, 1969
29. Prim RC: Shortest connection networks and some generalizations. *Bell Syst Techn J* 36:1389–1401, 1957
30. Sokal RR, Sneath PHA: *Principles of Numerical Taxonomy*. San Francisco, Freeman, 1963