PHYLOGENETIC ANALYSIS: MODELS AND ESTIMATION PROCEDURES[1]

L. L. CAVALLI-SFORZA AND A. W. F. EDWARDS[2]
International Laboratory of Genetics and Biophysics, Naples
Pavia Section, Istituto di Genetica, Università di Pavia, Italy

Received March 30, 1966

Acceptance of the theory of evolution as the means of explaining observed similarities and differences among organisms invites the construction of trees of descent purporting to show evolutionary relationships. Whether such trees are based on fossil or living specimens, they may often be criticized for having a high subjective element. The purpose of this paper is to show how suitable evolutionary models can be constructed and applied objectively. In it we amplify and extend the methods we have given in previous communications (Edwards and Cavalli-Sforza, 1963*a*, *b*, 1964, 1965; Cavalli-Sforza and Edwards, 1964, 1966; Cavalli-Sforza, Barrai and Edwards, 1964; Cavalli-Sforza, 1966).

Considering the great variety of information provided by living organisms, it is clear that the type of data will affect both the method of treatment and the validity of the results: the higher the correlation of data and genotype, the greater is the validity likely to be. Information on nucleic acid and protein structure comes first in the scale of relevance, and that on phenotypic measurements last; discrete and continuous variation demand different treatments, and evolutionary models appropriate to both cases will therefore be required for estimation purposes. Differences which are the result of mutation are formally discrete, and evolution at the molecular level thus needs discontinuous treatment; but even in this case the limit of observation may turn the data into the continuous type, as happens, for instance, when the similarity in nucleotide sequences in the nucleic acids of two organisms is measured by hybridization techniques, or when differences between closely-related organisms are examined. In the latter case the differences may be sufficiently small to suggest that the analysis be carried out at the level of gene frequencies, semi-continuous variables which may be treated as continuous. We will be especially concerned to develop the continuous treatment, the discontinuous one being more easily obtained, in a parallel manner.

In addition to the relevance of the data, the validity of the derived evolutionary tree will be strongly dependent on the correctness of the evolutionary model used as the basis for estimation. This will be limited, first because the dynamics of evolution is not fully understood, secondly because the values of some parameters (such as selective coefficients) are unknown, or known with low accuracy, and thirdly because there are mathematical and technical limitations as to how complex the model can be.

Although data suitable for our type of evolutionary study may seem to be largely taxonomic, it should be noted that the aim of this work is not the same as that of taxonomy, as the word is normally understood (see Edwards and Cavalli-Sforza, 1964); in particular, "numerical taxonomy" (Sokal and Sneath, 1963) is not primarily concerned with phylogeny, and the fact that the techniques to be described here and those of numerical taxonomy both involve the treatment of "taxonomic" data should not be allowed to mask the differ-

[1] This paper appeared in a recent supplement of the American Journal of Human Genetics.

[2] Present address: Department of Statistics, University of Aberdeen, U.K.

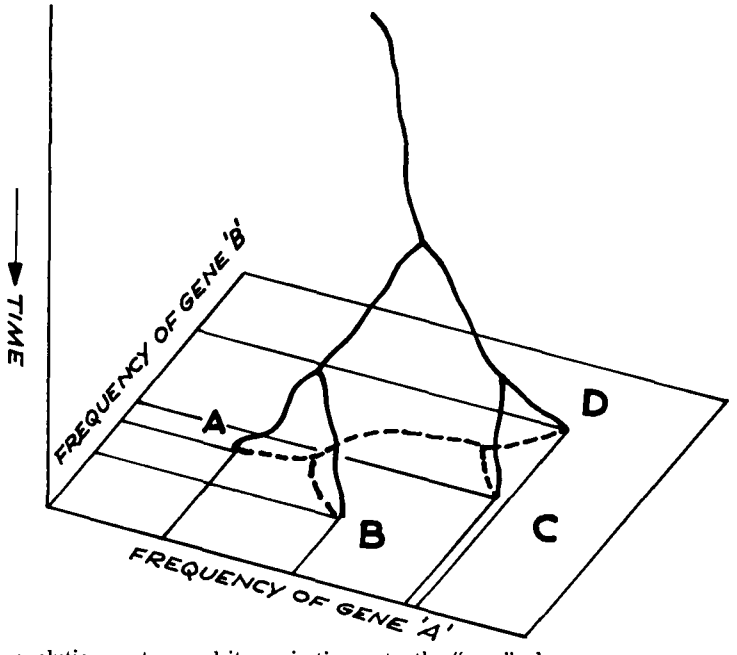


FIG. 1. An evolutionary tree and its projection onto the "now" plane.

ences between them, either at the logical or methodological levels.

EVOLUTION AS A BRANCHING PROCESS

Evolution can only be described in terms of the characters that are changing, and it is convenient to represent such changes in a multidimensional character-space in which each population occupies a position determined by the values of the characters it exhibits. In this paper the word "population" will be used to denote one of the group whose diversity is under study; it might refer to a species, a race, or even a single organism.

If a time dimension, everywhere normal to the character-space, is added, the course of evolution (were it but known) could be seen as a tree, whose branches split as populations diverge, unite as they hybridize, and end as they become extinct. Living populations would be represented by the intercept of the tree and the "now" plane (Fig. 1). In the case of discontinuous characters, the character-space would consist of a lattice of points, but to regard it as continuous will often be a good enough approximation, as indicated earlier. Data, such as gene frequencies or other measurements, will be represented by points in the space-time of Figure 1, and the problem of tracing evolutionary history will be that of fitting a suitable tree to these points. It may be noted, however, that information from the past is in practice available only for data which in other respects are less satisfactory than gene frequencies since their genetic basis is very imperfectly known, as is the case with osteometric data. But with gene frequencies data are only available, in general, for living populations, so that points will be restricted to the "now" plane. With such data we may only be able to reconstruct a projection of the tree onto the "now" plane (Fig. 1), in which case complete information on the position of the first split will not be preserved. Reconstruction of the tree in space and time will be possible, however, if we are willing to make hypotheses about the mode and speed of evolution.

Our genetic reasoning will be almost entirely confined to the analysis of gene frequencies among present-day populations, though it is clearly possible to extend it to other cases. In particular, once methods have been set up for estimating the course of evolution from present-day data, they can be extended without difficulty to include data from the past. Such an extension involves no logical jump and little increase in mathematical complexity.

The proper basis for the study of evolutionary divergence will be provided by the transformation of the space-time which makes a unit vector, in whatever direction (normal to time) and in whatever part of the space, correspond to the amount of evolutionary change expected in unit time. Such a transformed space-time will be homogeneous and isotropic with respect to evolutionary progress; in some problems it may not be Euclidean, or it may not even be possible to formulate the problem in geometrical terms, but the Euclidean representation is the simplest possible, and will suffice for the development of the argument. The correct transformation will, of course, depend on the evolutionary model and the type of data available, and in the case of gene frequencies will be treated below.

The Genetic Basis

Of the major evolutionary forces—mutation, migration, selection, and drift—we shall not incorporate the first two into our model. *Mutation pressure* is known to be usually very small compared with other pressures, so that its neglect, or its confounding with the other pressures, is reasonable: we are not here concerned with mutation in its role as the source of variation. *Migration* need not be considered at all if the evolving populations have differentiated past the specific stage; but this will not be true in most of the cases in which gene frequencies are useful, and its omission may be a source of error. We can, however, justify this practice: On the one hand, small migration rates will act essentially as almost-random disturbances partially buffering the variation due to random drift, so that the omission of migration of this magnitude is only apparent; on the other hand, large migration rates must appear only as rare accidents in a given evolutionary tree, and a "migrational accident," such as the fusion, partial or whole, of two populations, would give rise to a loop in Figure 1. The major difficulty in extending the analysis to include such loops is that they bring about an enormous increase in the number of possible trees to be examined. If, however, one or two loops are known, or assumed, to have taken place at given times and between given populations, their consideration may be practicable, although it will not be attempted in this paper.

Random genetic drift is the name given to the variation in gene frequencies which inevitably accompanies the formation of the next generation, depending, as it does, on a sample of genes from the former generation. Ignoring other sources of variation, in statistical terms this corresponds to a random walk of the gene frequencies in time. Whenever a population splits, the two branches will independently undergo random walks which will give rise to divergence between them, generally increasing with increasing time since branching. The rate of the random walk will depend on the size of the population in question (Wright's "effective breeding size") and the mating structure. The smaller the population, the greater the random variation, and thus the faster the random walk.

Under a suitable transformation this rate will be independent of the particular gene considered: the variance of a gene frequency p due to drift is approximately $p(1-p)\ (1-e^{-t/2N})$, where N is the effective size of the population and t the time elapsed in generations (Kimura, 1955). For estimation purposes, this can be made nearly independent of p by means of the angular transformation $p = \sin^2 \theta$, the variance of θ then being approximately

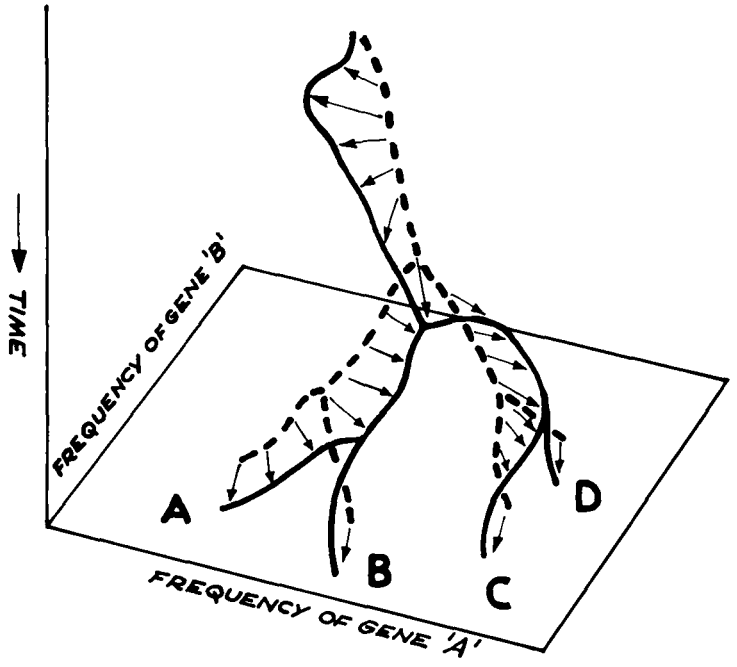


FIG. 2. The same tree as in Figure 1, but with the addition of directional selection constant in space but not time, indicated by arrows. The evolutionary paths without selection (broken lines) are translated into new paths (solid lines) without affecting the relative positions of the populations.

$(1 - e^{-t/2N})/4$ (see Fisher, 1958). When $2N$ is large compared with t, this variance reduces nearly to $t/8N$.

In recent papers (Cavalli-Sforza, Barrai and Edwards, 1964; Cavalli-Sforza, 1966) we have given some reasons why we believe that random drift is likely to be more important than formerly believed in determining the variation in gene frequencies in man, and similar considerations may be applied to other organisms.

Selection may be constant in space and time, or may vary in one or the other or both. If selection is constant in space it gives rise to a shift in the gene frequencies of all the populations studied (directional selection, Fig. 2). It would be detected as a trend in the gene frequencies if data from the past were available, but it is not, by definition, a cause of differentiation between populations, and is not detectable in data solely from present-day populations.

Variable selection, however, is probably a major factor in causing divergence. If the variation is sufficiently rapid and haphazard, Kimura's (1954) model of "selective drift" will be appropriate. This model gives rise to a variation in gene frequencies which is almost indistinguishable from the effects of random genetic drift, apart from the behavior at extreme gene frequencies (0 or 1). Thus the consequence of "selective drift" will also be a fluctuation of gene frequencies akin to a random walk, with rate depending on the variability of the selective coefficients. The intensity of this variation will be a property of the gene concerned. If selection is of the stabilizing kind, such as in heterozygotic advantage, it will cause a reduction in the variance of gene frequencies, and its effect will be more or less inextricably confounded with that of random drift.

Prolonged periods of selection peculiar to individual populations will not be de-

tectable without data from the past, and no method of phylogenetic analysis can alter the fact that any observed diversity can be explained by any evolutionary tree provided we are willing to postulate the necessary selection. Our methods invoke no such specific postulates, although if other evidence on selection in a particular population is available it should of course be taken into account. As we shall see, there are not enough degrees of freedom to estimate selection specific to individual populations.

To sum up, selective trends will be detectable only if data from the past are available. Random drift will create a random walk of the gene frequencies, to which selective drift will add, and stabilizing selection subtract, speed. If major and sudden shifts in gene frequency have occurred, because of large selective accidents or other bottlenecks in population development, they will be a source of inaccuracy in the analysis, although their presence may be detectable as departures from expectation indicated by the goodness-of-fit of the models we have developed. The whole problem of the robustness of our methods with respect to varying forms of selection, migration, and other departures from the simple model, may best be studied by Monte Carlo methods, trees being generated according to specific hypotheses, and the estimated forms compared with the known ones.

The Statistical Basis

We have given reasons above why the variation in gene frequencies of each population may be represented by a random walk. To keep the model as simple as possible, it will be supposed that no population becomes extinct, that at each split the daughter populations are both identical to their parent, and that each population is independent of every other one. We shall now examine the statistical assumptions underlying the methods of analysis.

We need hypotheses on the mode of splitting of populations and on the properties of the random walks in the individual branches of the evolutionary tree. The first type of hypothesis will affect only the form of the tree, while the second will affect its dimensions as well; both form and dimensions requiring to be estimated. As we shall see, the number of forms increases very fast with the number of populations. Depending on this number, we shall either test all possible forms or limit our analysis to a group of promising ones.

Given a particular form, the optimum tree may perhaps be estimated by maximum likelihood, or some other method, and the choice between forms will then depend on a comparison of their likelihoods or other appropriate criteria. The simplest model for splitting, which we shall use here, is that in which it occurs at random, as in a Yule process (Yule, 1924): When there are n populations each is assumed to have equal probability, in a given time interval, of generating the $(n+1)$th. On this basis the probabilities of the different forms may be written down explicitly when n is fairly small, and such probabilities will be used, where possible, as prior probabilities in the estimation procedure. The technique for enumeration is to generate the possible "topologies" (that is, tree forms irrespective of the placings of the populations on the terminal branches) for n populations from those for $n-1$ by allowing each terminal branch to split in turn, keeping track of the probabilities. The labelled populations are then distributed on the terminal branches in all possible distinct ways for each topology, each arrangement having equal probability for a given topology, and the final probability of each form thus calculated. This procedure will be clarified in the example to be given below. The calculation of the probabilities for large n has been considered by Harding (1967).

The properties to be assumed for the random walk depend directly on the biological assumptions. Using the transformation of the gene frequencies given above (which will later be generalized to the

multi-allelic case), the distribution of the transformed variates will be approximately Gaussian with variance proportional to the time elapsed, while the mean will be constant in the absence of directional selection. We have already decided not to consider individual directional selection of a prolonged nature, and directional selection which is the same for all populations at any given time will not affect their relative movements, as has been indicated above, and therefore need not be considered, although it must be remembered that any inference about the gene frequencies in an earlier population will be based on its presumed absence. Thus the random walk of the gene frequencies may be regarded as a Brownian-motion or Wiener process in the transformed space.

Various models of increasing complexity can be imagined which could, under ideal conditions, be fitted in succession to the data, stopping at the lowest level necessary for a good fit:

1) The simplest model is one in which the Brownian motion has a constant rate which is the same for all characters at all times. This represents the case in which random drift alone determines the variation in gene frequencies, and population size and structure are taken to be constant.

2) The Brownian motion does not have the same rate for all characters. A simple transformation of the scale of each character to standardize the variances will allow us to include selective drift among the evolutionary forces considered, in addition to random drift, since selective drift causes the variances of characters to differ one from another. Stabilizing selection will be confounded with selective drift and thus also included. As an improvement on simply transforming the scales, the variances can themselves be estimated simultaneously with the estimation of the tree.

3) The Brownian motion is not constant in time. If it varied randomly, owing to random fluctuations in population size, this could be taken into account.

4) If nothing is known about the probability distributions of selective coefficients or population sizes, we may restrict the assumptions to that of the independence of events in individual populations, thus retaining the idea of Brownian motion due to the combined effect of random and selective drift, but with no knowledge of local variations in its rate.

We have worked out three essentially different methods. The first is estimation based on the method of maximum likelihood applied to specific models; we have fully developed this approach for model (1) and in outline for model (2), though, of the two, only the former will be described in this paper. Model (3) awaits treatment by maximum likelihood, but model (4) cannot be handled in this way owing to the lack of information on the probability distributions of selective coefficients and population sizes. Indeed, as has been mentioned above, model (4) cannot be treated rigorously by any method.

The second method is the "method of minimum evolution" (Edwards and Cavalli-Sforza, 1963*a*), which uses the intuitive idea that a plausible estimate of the projection of the evolutionary tree onto the "now" plane is given by that tree which invokes the minimum total amount of evolution (see Fig. 1). The assumptions underlying this method are not too clear; it may go some way towards handling model (4), but its success is probably due to the closeness of the solution it gives to the projection of the "maximum-likelihood" tree. The extent of the similarity merits further investigation, and experience with simulated trees should clarify its logical status. It certainly cannot be justified on the grounds that evolution *proceeds* according to some minimum principle, as recently suggested by Camin and Sokal (1965) when they applied it to the discrete case. Its success with discrete data (see also Zuckerkandl, 1965) must also be attributed to its closeness to the solution which a proper probability model would give.

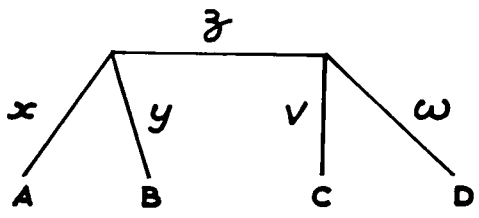


FIG. 3. Model for the "additive" tree, showing expected branch lengths.

Our third method may be called that of the "additive tree" (Cavalli-Sforza and Edwards, 1964). It assumes that distances along the tree are additive, thus implying independence of evolution in all the branches. For example, in Figure 3, the evolutionary distance between population A and the last common ancestor of populations A and B is called x, and that between population B and the same ancestor, y; it is then assumed that the observed distance between A and B, d_{AB}, is an estimate of $x+y$. For this tree we can thus set up six equations:

$$d_{AB} = x + y + \text{"error"}$$
$$d_{AC} = x + z + v + \text{"error"}$$
$$d_{AD} = x + z + w + \text{"error"}$$
$$d_{BC} = y + z + v + \text{"error"}$$
$$d_{BD} = y + z + w + \text{"error"}$$
$$\text{and} \quad d_{CD} = v + w + \text{"error."}$$

The "error" terms represent the departures of the observed from the expected distances, and a method of estimation of v, w, x, y, and z is to minimize the sum of the squared errors—the method of least squares.

There is no trunk to the tree of Figure 3 because it is impossible to obtain information on the first split, as in the method of minimum evolution, and the branch corresponding to z is therefore represented by a single line. The tree obtained in this way cannot be represented in the character space, as may be seen by considering the case of three populations, each a distance d from the other. The three-branched least-squares tree (which has zero residual sum of squares) has branches each of length $d/2$, which do not meet if they are supposed to extend inwards from the vertices of an equilateral triangle of side d. They could be made to meet by imposing the appropriate restriction, but this would upset the simplicity of the estimation procedure, and has not been tried. The method does not, therefore, estimate the positions of nodes, either in time or space; but it has the advantage that it can be applied to distances which cannot be represented in a metric space. Apart from the assumption of independence in the separate branches, the justification of this method seems to be much the same as that of the method of minimum evolution, considered above.

DETAILS OF THE METHODOLOGY

Number of tree forms.—It has been noted above that the number of tree forms increases very rapidly with increasing n, the number of populations. The number of forms is in fact $3.5.7 \ldots (2n-3) = (2n-3)!/[2^{n-2}(n-2)!]$ when the first split can be recognized. Such a tree has $2n-1$ branches (including the trunk) so that, in progressing from a tree with $n-1$ populations to one with n, the new branch may be inserted in any one of $2(n-1)-1 = 2n-3$ places. At $n=10$ this number is 34,459,425. When there is no information about the first split, there is no trunk, and the number of forms is reduced to $3.5.7 \ldots (2n-5)$, giving 2,027,025 for $n=10$.

It is interesting to note, in passing, that the number of "topologies" (tree forms irrespective of the placings of the populations on the terminal branches) is given, for trees with trunks, by a_n, where

$$a_n = a_1 a_{n-1} + a_2 a_{n-2} + \ldots + a_{(n-1)/2} a_{(n+1)/2} \qquad (n \text{ odd})$$

$$\text{and} \quad a_n = a_1 a_{n-1} + a_2 a_{n-2} + \ldots + \tfrac{1}{2} a_{n/2}(a_{n/2}+1) \qquad (n \text{ even}).$$

This solution is derived by considering the number of ways a tree can be constructed by uniting smaller trees at their trunks,

and follows Polya (1937); similar problems had been considered by Cayley (1857, 1859) in connection with the number of ways brackets can be inserted in an algebraic expression. We find $a_{10} = 98$.

With such large numbers it is thus important, when there are more than seven or eight populations, to have methods of generating promising forms for estimation. We use several somewhat intuitive methods (in the hope that they corroborate each other), and then use the forms so found, and similar ones, as a basis for metric estimation, the final choice being according to whatever criterion, such as likelihood, is appropriate.

Choice of promising forms.—Intuitively it is reasonable to suppose that present-day populations near to each other in the character space should be clustered together on the same part of the evolutionary tree, so that methods of clustering points should give some insight into the most promising form. Among the methods used, one (Edwards and Cavalli-Sforza, 1963*b*, 1965; see also Ward, 1963) divides the populations into the two clusters for which the between-clusters sum of squares is a maximum, and each cluster is then similarly treated, and so on until a complete breakdown of the original cluster has been made.

Apart from this method, we have used two others especially well-suited to the application of the method of minimum evolution. The first is based on a theorem due to Prim (1957), who showed that when branches are constrained to meet only at populations, the net of shortest length may be found by listing all the pairwise distances between populations in order of ascending magnitude and allocating branches successively to these distances, omitting any branch which completes a loop. The resulting net is the shortest net conditional on each node coinciding with one of the populations; removing the constraint, the net may be shortened step by step, as described below. The second method uses the same iterative procedure, but starting from the case in which all the nodes coincide at a single point, so chosen that the total length is then a minimum.

In the absence of a means of maximizing likelihood over different tree forms, these methods have shown promise of leading to near-optimal forms. The last two do not, however, give any indication of the position of the first split, so that several positions must be tried.

Transformation of data.—We will concentrate on populations of nonhaploid organisms, whose evolution is suitably studied by way of continuous data, be they gene frequencies or metrical characters. We have mentioned the usefulness of the angular transformation of gene frequencies, whereby the variances are standardized irrespective of the frequencies themselves (at least in the interval 0.05 to 0.95). It was pointed out by Fisher (personal communication; see also Cavalli-Sforza and Conterio, 1959) that this transformation could be generalized to multi-allelic loci, as follows.

If each of m alleles at the locus is allotted a Cartesian axis in a Euclidean space of m dimensions, and a population with gene frequencies $p_1, p_2, \ldots p_m$ is represented by the vector $(\sqrt{p_1}\ \sqrt{p_2} \ldots \sqrt{p_m})$, then the space of possible populations is the $\frac{1}{2}^m$th part of the surface of the unit hypersphere in which all the coordinates are non-negative, the population being represented by a point unit distance from the origin with direction cosines given by the above vector (Fig. 4). It follows that the angular distance between two populations with gene frequencies $p_1, p_2, \ldots p_m$ and $p_1', p_2', \ldots p_m'$ is given by θ where $\cos\theta = \sum_{i=1}^{m} \sqrt{p_i p_i'}$. Since $\theta = \pi/2$ corresponds to a complete gene substitution, it is convenient to work in terms of $2\theta/\pi$, where θ is in radians, for the unit distance is then one gene substitution.

In this representation a population may be thought of as pursuing a random walk in the curved space. Since this space is finite, with known bounds, the Gaussian

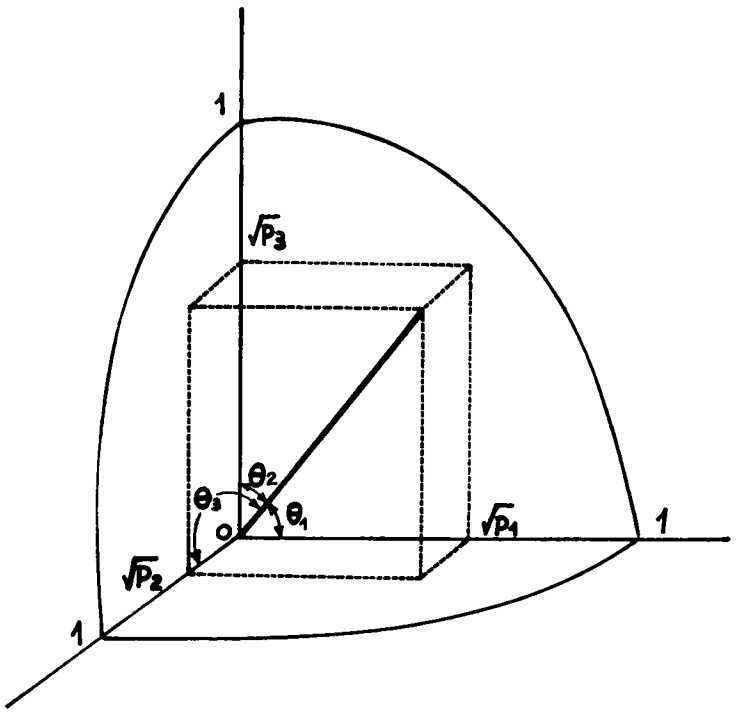


FIG. 4. Representation of a population, with gene frequencies p_1, p_2, p_3 at a single triallelic locus, on the octant of a sphere.

approximation to the gene frequency distribution will only hold if the variance is sufficiently small and the population sufficiently far removed from an edge of the space for edge effects to be neglected. The method of maximum likelihood could ideally now be applied to a model based on this transformation of the data, the log-likelihoods being summed over loci, but it turns out to be intractable owing to the curved space and difficulties with the coordinate system, so that it is necessary to approximate the curved space in the region of the populations by a Euclidean space of $(m-1)$ dimensions by means of a projection of one onto the other. An orthogonal projection onto the hyperplane tangent to the hypersphere at the centroid of the populations should suffice, although in the present work we have simply used, as the distance between two populations an arc $2\theta/\pi$ apart, the length of the chord joining them, which is $(2\sqrt{2}/\pi)$ $(\sqrt{1-\cos\theta})$. Thus the m-dimensional Euclidean space in which the hypersphere is embedded has itself been employed.

These Euclidean spaces for the separate loci (assumed unlinked) may then be combined, distances being given by Pythagoras' theorem in many dimensions, so that the square of the distance between two populations is given by the sum of the squared distances for each locus. In this way the data are represented in a Euclidean space, the scale of which is one unit per gene substitution.

Another type of continuous data of some interest is that in which measurements can only be made directly on the pairwise distances between populations. Such is the case, for instance, when "immunological" distances between populations are investigated by serological methods, or when differences in nucleotide sequences are estimated using hybridization procedures with nucleic acids. In these cases, data consist

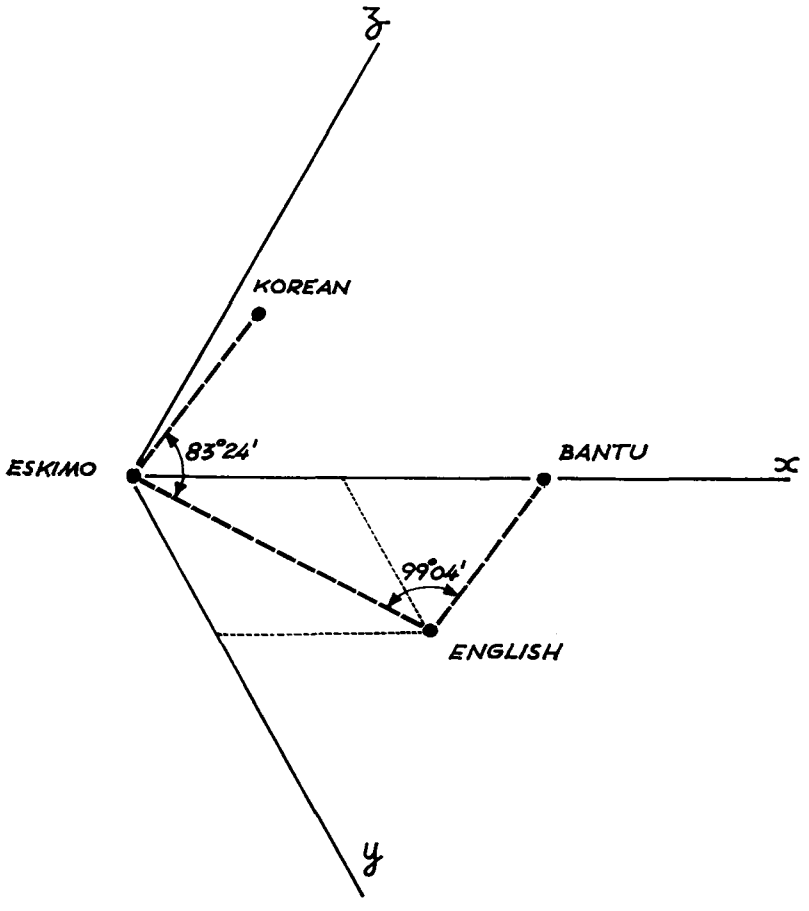


FIG. 5. Representation of the four populations in Euclidean space, showing the Prim network.

of a triangular matrix of the pairwise distances between populations, which is also the form to which multi-locus gene frequency data has been reduced by the methods described above. But whereas in the latter case the erection of a Cartesian coordinate system in Euclidean space by repeated applications of Pythagoras' theorem is bound to succeed, in the former case it will often fail, the method then generating complex numbers. But even if a nonlinear scale transformation allowing Euclidean representation cannot be found, the procedures for cluster analysis and for finding an "additive tree" by least squares may still be applied. One advantageous by-product of generating coordinates from the pairwise distances is that the maximum number of dimensions required is one less than the number of populations, however many characters (and hence dimensions) there were originally.

The case of ordinary metrical characters has been treated only to a very limited extent because of the difficulties in interpreting their genetic basis. We considered them when analyzing anthropometric data in one of the earlier papers (Cavalli-Sforza and Edwards, 1964), but the results were not very encouraging. It seems that the best procedure is to transform the original characters into a set of uncorrelated standardized variables, using the within-populations dispersion matrix (see, for example,

Rao, 1952). The lack of the complete matrix for the data just mentioned must have contributed to the unsatisfactory nature of the solution.

Estimation based on maximum likelihood.—In Euclidean space of p dimensions, the probability density a distance d from an original population after a time t has elapsed is $(1/\sigma\sqrt{2\pi t})^p \exp(-d^2/2t\sigma^2)$, owing to the Gaussian nature of the random walk, where σ^2 is the variance per unit time. The log-likelihood of a branch of length d and interval t is therefore $-(d^2/2t\sigma^2 + \frac{1}{2}p.\log 2t\sigma^2 + \frac{1}{2}p.\log \pi)$. Writing T for $2t\sigma^2$ (time and variance being confounded in their product, as is to be expected), omitting the constant and changing the sign, this becomes $d^2/T + \frac{1}{2}p.\log T$.

Let x_{ir} be the r^{th} spatial coordinate of the i^{th} node (or population) and t_i its time coordinate, measured backwards from the "now" plane. For the branch joining nodes i and j $(t_j > t_i)$ the above expression then becomes

$$\sum_r (x_{jr} - x_{ir})^2/(t_j - t_i) + \tfrac{1}{2}p.\log (t_j - t_i). \quad (1)$$

Summing over all the branches (denoted by (i, j)), the quantity to be minimized is

$$L = \sum_{(i,j)} \left[\sum_r (x_{jr} - x_{ir})^2/(t_j - t_i)\right] + \\ + \tfrac{1}{2}p \sum_{(i,j)} \log (t_j - t_i). \quad (2)$$

If j, k, and l are the three nodes (or, in the case of k and l, possibly populations) connected by single branches to node i $(t_j > t_i > t_k, t_l)$, then

$$\tfrac{1}{2}\frac{\partial L}{\partial x_{ir}} = \frac{x_{ir} - x_{jr}}{t_j - t_i} + \frac{x_{ir} - x_{kr}}{t_i - t_k} + \frac{x_{ir} - x_{lr}}{t_i - t_l} \\ = 0. \quad (3)$$

Each node yields a similar equation linear in the $x_{.r}$, the net result being a system of linear equations which enables the r^{th} spatial coordinates of the nodes to be expressed in terms of the r^{th} spatial coordinates of the populations and the time coordinates of the nodes, by inverting a matrix which is the same for each dimension. Assuming initial values for the time coordinates, the numerical values thus found can be inserted in the first and second partial derivatives of L with respect to the t's (these will not be quoted explicitly, but may be immediately written down from (2)), giving the scores and the information matrix. Corrections to the t's are found in the usual way, and the cycle repeated. By means of this reduction of the likelihood equations the order of the information matrix shrinks from $(n-1)(p+1)$ to $(n-1)$.

In order to begin iteration, initial estimates of the t's are required, and these may be found from cluster analysis. Considering two populations at time $T = 0$ a distance d apart, the maximum-likelihood estimate of the time coordinate of the subtending node is $d^2/2p$, while the variance of the two-population cluster is $d^2/4$. Putting the time coordinate of the subtending node equal to $2/p$ times the variance of the cluster, of however many populations, thus provides a rough estimate.

It may be noted that of the $n(p+1)$ original degrees of freedom $(n-1)(p+1)$ are used in estimating the parameters, leaving $(p+1)$. If it is not assumed that the variance of the random walks is the same in each direction, $(p-1)$ relative variances must also be estimated, leaving two degrees of freedom for a goodness-of-fit test. There are not sufficient degrees of freedom to estimate covariances as well, although these will be small if the original data have been transformed into uncorrelated variables using the between-populations dispersion matrix.

Unfortunately, except in the very simplest cases, this straightforward application of the method of maximum likelihood leads into difficulties due to the fact that the likelihood surface contains singularities, as may be seen from the following example (Cavalli-Sforza and Edwards, 1966). Since the log-likelihood of the tree is found by summing the log-likelihoods of the individual branches, each node should, in the final

solution, be in the maximum-likelihood position with respect to the three adjacent nodes if these are regarded as fixed. In particular, given the positions and times of the second and third splits, those of the first should be determinable. Let us therefore consider the simple case of the determination of the origin of a Brownian-motion process when we have observed just two different populations: at time t_1 population 1 was at x_1, and at time t_2 population 2 was at x_2. Let the origin of the process, which we are required to estimate, be at (X, T). The quantity to be minimized is, from equation (2) above,

$$L = \frac{(X-x_1)^2}{T-t_1} + \frac{(X-x_2)^2}{T-t_2} + \\ + \tfrac{1}{2}\log(T-t_1) + \tfrac{1}{2}\log(T-t_2). \quad (4)$$

This surface can be plotted for varying X and T, but it is already obvious that it contains an unfortunate characteristic: suppose $t_1 > t_2$ (so that population 1 is earlier in time, which, it may be remembered, is being measured backwards), and suppose that we trace the path $X - x_1 = T - t_1$ along the surface towards (x_1, t_1). Then the first term in L will tend to zero, the second will tend to the constant $(x_1 - x_2)^2/(t_1 - t_2)$, but the third will become indefinitely large and negative as T approaches t_1, although the fourth will tend to $\frac{1}{2}\log(t_1 - t_2)$. Thus, provided the two observations were not made simultaneously, in which case $t_1 = t_2$, the negative log-likelihood can be made as large and negative as we please, and hence the log-likelihood as large as we please, simply by letting the origin approach the first observation along a certain path; the surface contains, in fact, a singularity at the point (x_1, t_1), and any iterative procedure, unless it is started near a well-behaved peak elsewhere in the surface, will lead to the coincidence of the origin with the first observation. It is convergence to such singularities which has led to the trivial solutions obtained when studying complete trees. Needless to say, such solutions are intuitively unacceptable, although this is not the place to consider in detail the problems of inference which they expose. Short of portraying the entire likelihood surface, which is impossible owing to the number of parameters involved, there is no fully satisfactory solution.

However, we have developed another approach which seems to us to be moderately satisfactory. Recalling that, given the time coordinates, the spatial coordinates of the nodes can be found by maximum likelihood without difficulty, the problem can be solved by finding the time coordinates by another method. One such method has already been given above, but it was intended only as an initial approximation. We now propose to fit, according to the least-squares criterion, the observed length squared of each branch to its expected squared length derived from the interval between its ends. On the Gaussian model this expected squared length is easily seen to be the interval itself.

We therefore minimize, in the notation of equations 1 to 3,

$$S = \sum_{(i,j)} [\sum_r (x_{jr} - x_{ir})^2 - (t_j - t_i)]^2, \quad (5)$$

with respect to the t_i, the time coordinates of the nodes. Differentiating with respect to t_i,

$$\tfrac{1}{2}\frac{\partial S}{\partial t_i} = \sum_r (x_{jr} - x_{ir})^2 - \sum_r (x_{ir} - x_{kr})^2 - \\ - \sum_r (x_{ir} - x_{lr})^2 + 3t_i - t_j - t_k - t_l \\ = 0. \quad (6)$$

Each node yields a similar equation linear in the t_i, enabling them to be expressed in terms of the x_{ir} by inverting a matrix. The values thus obtained are used in the method of maximum likelihood to generate new spatial coordinates for the nodes (equation 3 above), and the cycle repeated until convergence is attained. The procedure amounts to the iterative solution of two sets of equations exemplified by 3 and 6 above, and is a generalization of that pro-

TABLE 1. *Blood-group gene frequencies characterizing four populations.*

	Eskimo	Bantu	English	Korean
A_1	0.2914	0.1034	0.2090	0.2208
A_2	0.0	0.0866	0.0696	0.0
B	0.0316	0.1200	0.0612	0.2069
O	0.6770	0.6900	0.6602	0.5723
CDE	0.0	0.0	0.0024	0.0082
CDe	0.4985	0.1400	0.4205	0.6197
cDE	0.4906	0.0100	0.1411	0.3148
cDe	0.0109	0.6000	0.0257	0.0573
Cde	0.0	0.0200	0.0098	0.0
cdE	0.0	0.0	0.0119	0.0
cde	0.0	0.2300	0.3886	0.0
MS	0.1719	0.0900	0.2377	0.0245
Ms	0.6703	0.4800	0.3048	0.4615
NS	0.0	0.0400	0.0703	0.0646
Ns	0.1578	0.3900	0.3872	0.4494
Fy^a	0.7500	0.0600	0.4213	0.9950
Fy^b	0.2500	0.9400	0.5787	0.0050
Di^a	0.0	0.0	0.0	0.0313
Di^b	1.0	1.0	1.0	0.9687

posed by Cavalli-Sforza and Edwards (1966) to solve the two-population example considered earlier.

Having had to invoke the method of least-squares for reasons of expediency, it is necessary to defend the decision not to solve the problem entirely by this method (maximizing S with respect to the spatial as well as the time coordinates of the nodes), but rather to rely on a hybrid between least-squares and maximum-likelihood. We retain the method of maximum likelihood because

1) We prefer it as a method of estimation where it does not break down;

2) It is more tractable in the present case, as may be seen by differentiating equation 5 with respect to the $x_{..}$;

3) Likelihood provides a more useful criterion of acceptability than the residual sum of squares, as it may be directly compounded with prior probabilities.

Although it is very probable, it remains to be shown formally that the least-squares solution converges to the local maximum of the likelihood in well-behaved examples where the singularities may be discarded, and it will be worthwhile to try the full maximum-likelihood procedure using the estimates obtained by the above procedure as starting values.

It has been pointed out to us by D. G. Kendall and D. F. Kerridge (separate personal communications) that, in writing down the likelihood of a tree, we have omitted the contribution due to the "Poisson" nature of the Yule process. If we knew that n, and only n, populations *could* have been produced by the process, then the likelihood surface for the times of the splits would indeed be uniform, as we have supposed. But this is not really our model: In fact, by using a Yule process, we admit the possibility that numbers other than n could have been generated on the same model. The likelihood surface, conditional on n, is then easily seen to be proportional to $e^{-\lambda \Sigma t_i}$, where Σt_i is the sum of the times of the nodes and λ is the rate of the process. The question arises, however, of whether even this is the appropriate model, for we have not in fact observed a tree with exactly n populations, but rather we have chosen to observe n from a tree which may have had more. The likelihood surface for the nodes of a tree with n populations which have been chosen at random from a randomly-generated tree with m $(m \geqslant n)$ populations appears to be unknown. It may even be uniform.

However, even the use of the likelihood $e^{-\lambda \Sigma t_i}$ would not affect our argument, for it would weight the likelihood in favor of still shorter times for the branch lengths, thus aggravating rather than alleviating the problem of the singularities. Recourse to the least-squares modification would still be necessary, and, since the likelihood is only used to derive the spatial coordinates for given times, the addition of the quantity $-\lambda \Sigma t_i$ to the log-likelihood would not affect the estimates. It would, of course, affect the likelihood comparisons of different tree forms, though since we do not know λ, and cannot estimate it by maximum likelihood, and since we are in

TABLE 2. *Gene substitutions separating four populations, measured along the arc ($2\theta/\pi$) and along the chord ($[2\sqrt{2}/\pi]\sqrt{1-\cos\theta}$).*

		A_1A_2BO	Rh	MNSs	Fy	Di	Combined
Eskimo	arc	0.2585	0.7274	0.2222	0.5091	0.0	–
—							
Bantu	chord	0.2568	0.6886	0.2211	0.4957	0.0	0.9136
Eskimo	arc	0.1822	0.4718	0.2852	0.2170	0.0	–
—							
English	chord	0.1816	0.4611	0.2828	0.2159	0.0	0.6101
Eskimo	arc	0.1874	0.1461	0.3050	0.2883	0.1132	–
—							
Korean	chord	0.1868	0.1457	0.3021	0.2858	0.1131	0.4918
Bantu	arc	0.1094	0.4907	0.1554	0.2922	0.0	–
—							
English	chord	0.1092	0.4787	0.1549	0.2896	0.0	0.5907
Bantu	arc	0.2270	0.6398	0.1024	0.7974	0.1132	–
—							
Korean	chord	0.2258	0.6132	0.1023	0.6871	0.1131	1.0036
English	arc	0.2178	0.4465	0.2293	0.5052	0.1132	–
—							
Korean	chord	0.2167	0.4374	0.2280	0.4921	0.1131	0.7384

any case not certain of the model, we have omitted it in the following example.

It should be noted that, in the foregoing argument, if the t_i are in units of $2\sigma^2$ according to our standard notation, λ is in units of $1/2\sigma^2$, so that it is the expression $\lambda/2\sigma^2$ that is really relevant.

Estimation by the method of minimum evolution.—The method of minimum evolution (Edwards and Cavalli-Sforza, 1963*a*) uses the intuitive idea that a plausible estimate of the projection of the evolutionary tree onto the "now" plane is given by that tree uniting all the populations which invokes the minimum total amount of evolution. With suitably transformed scales, a unit quantity of evolution is a unit vector in this space, and therefore the required tree is that with the minimum total length possible. As has been mentioned, the position of the first split is undefined in the projection, and, of course, the method of minimum evolution gives no information on the time coordinates of the nodes.

The problem of finding the minimum-length tree may be referred to as the Steiner problem in p dimensions (see Courant and Robbins, 1960), and no algorism for its general solution is available, although various things are known. For example, it may be proved that, in any number of dimensions, each node must be the meet of just three branches mutually inclined at 120°, unless the node coincides with a population. It follows that, for a given tree form, a computer program may be written which will converge to the Steiner tree *for that form*, because if any two intersecting branches meet at an angle of less than 120° they may be replaced by a Y-shaped "Steiner triplet" whose inter-branch angles are 120°, the total length thus being shortened. Unless the tree form is close to the optimum, however, many branches will converge to zero length, but this fact may be used to give an indication of how the form can be improved. A Steiner tree with no branches of zero length (except those generated by nodes coinciding with populations) may be called "stable." There will usually be many stable nets for a given configuration of populations, so that stability itself is no guarantee that the minimum net has been

TABLE 3. *Coordinates of the four populations.*

	x	y	z
Eskimo	0.0	0.0	0.0
Bantu	0.9136	0.0	0.0
English	0.4695	0.3895	0.0
Korean	0.0379	0.0428	0.4885

found, and, as with other methods, many different forms must be investigated.

As indicated on page 557 above, we can try the form derived from Prim's network by calculating all the interbranch angles, and replacing the two branches subtending the smallest angle by the Steiner triplet, and so on, treating the angles in order of increasing size. Alternatively, we can proceed in the same way starting with all the nodes coinciding at a central point so chosen that the total length is then a minimum.

Since the method seeks the tree of minimum length, this length is the criterion of acceptability, akin to likelihood on the full model, but the relation between these two quantities needs further study. A more detailed account of these methods for solving Steiner's problem in p dimensions is given by Edwards (1967).

Estimation on the additive tree model. —If $\mathbf{D}$ is the column vector of the $\frac{1}{2}n \cdot (n-1)$ pairwise distances between populations, $\mathbf{W}$ the column vector of the $2n-2$ branch lengths to be estimated, and $\mathbf{B}$ the $\frac{1}{2}n(n-1) \times (2n-2)$ matrix specifying the form of the tree, in which the element in the r^{th} row and c^{th} column is 1 if the path uniting the r^{th} pair of populations includes the c^{th} branch, and 0 otherwise, then the least-squares estimate of $\mathbf{W}$ is given by $\hat{\mathbf{W}} = (\mathbf{B}^*\mathbf{B})^{-1}\mathbf{B}^*\mathbf{D}$, where $\mathbf{B}^*$ is the transpose of $\mathbf{B}$. Thus the corresponding expected value of $\mathbf{D}$ is $\hat{\mathbf{D}} = \mathbf{B}\hat{\mathbf{W}}$, the residual sum of squares is $\mathbf{D}^* \cdot (\mathbf{D} - \hat{\mathbf{D}})$ and it follows that the variance-covariance matrix of $\mathbf{W}$ is $\mathbf{D}^*(\mathbf{D} - \hat{\mathbf{D}}) \cdot (\mathbf{B}^*\mathbf{B})^{-1}/[\frac{1}{2}(n-1)(n-4)]$. If the errors in the observed pairwise distances are not independently distributed, or have unequal variances, the corresponding variance-covariance matrix should be incorporated in the estimation procedure, as described in Kendall and Stuart (1961); but this improvement has not been used in the present work.

TABLE 4. *Possible tree forms with four populations.*

(a) rooted

1	Esk	Ban	Eng	Kor
2	Esk	Ban	Kor	Eng
3	Esk	Eng	Ban	Kor
4	Esk	Eng	Kor	Ban
5	Esk	Kor	Ban	Eng
6	Esk	Kor	Eng	Ban
7	Ban	Eng	Esk	Kor
8	Ban	Eng	Kor	Esk
9	Ban	Kor	Esk	Eng
10	Ban	Kor	Eng	Esk
11	Eng	Kor	Esk	Ban
12	Eng	Kor	Ban	Esk

(each with probability 1/18)

13	Esk	Ban	Eng	Kor
14	Esk	Eng	Ban	Kor
15	Esk	Kor	Ban	Eng

(each with probability 1/9)

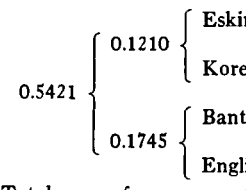
(b) unrooted

1′	Esk	Ban	Eng	Kor
2′	Esk	Eng	Ban	Kor
3′	Esk	Kor	Ban	Eng

(each with probability 1/3)

Corresponding to a branch of zero length in the method of minimum evolution, a branch of negative length in the least-squares method (there is no restriction against negative branches appearing) indicates the need to change the tree form. Indeed, it is not clear that a tree with no negative branches may always be found. The residual sum of squares is the criterion for choosing between trees of different form, although the introduction of prior probabilities seems to be impossible. In comparing branch lengths found by the "additive-tree" and "minimum-evolution" methods, it should be noted that not only

TABLE 5. *Best tree found by cluster analysis, showing sums of squares removed by each split.*



Total sum of squares = 0.8375

TABLE 6. *"Additive-tree" and "minimum-evolution" results for the unrooted tree forms.*

Form	Residual sum of squares ("additive tree")	Total amount of evolution ("minimum evolution")
1′	0.07054	1.7998
2′	0.08105	1.7850
3′	0.00037	1.6173

does the former method give trees unrepresentable in the character space, as mentioned earlier, but also that the longer branches are likely to be considerably shorter than their "minimum-evolution" counterparts.

AN EXAMPLE WITH FOUR POPULATIONS

In order to illustrate the use of our methods in detail, in this section we present an exhaustive treatment of a case with four populations.

The data consist of the gene frequencies for the five blood-group systems A_1A_2BO, RH (four sera), MNSs, Fy, and Di, for samples from four human populations: Eskimo, Bantu, English, and Korean. These data, which are given in Table 1, are part of a larger collection that we have used in previous communications. They are drawn from the following sources: Eskimo—Chown and Lewis (1959); Bantu—quoted by Zoutendyk, Kopeć, and Mourant (1955), with the subdivision of group *A* provided by Mourant (personal communication), and Di^a presumed absent; English—Race and Sanger (1959); Korean—Won, Shin, Kim, Swanson, and Matson (1960), with subdivision of MN by *S* and *s* according to the Japanese proportions reported by Lewis, Kaita, and Chown (1957). No special significance should be attached to this choice of samples, which was made largely for convenience.

From these data the values of $2\theta/\pi$ for the pairwise distances are calculated for each locus, and are given in Table 2. Thence the pairwise chords are found $[(2\sqrt{2}/\pi)(\sqrt{1-\cos\theta})]$, and the overall pairwise distances computed by taking the square root of the squared distances summed over loci, for each pair (Table 2). Finally (as far as the preparation of the data is concerned) a coordinate system, relative to arbitrary Cartesian axes, is erected for the four points, the coordinates being given in Table 3. The consequent arrangement of the points is shown in Figure 5.

Applying the appropriate formulae, we find there to be $5 \times 3 = 15$ possible rooted tree forms, and 3 possible forms without a root; these are listed in Table 4, together with their prior probabilities, which have been found by direct enumeration. In order to demonstrate the normal course of the analysis we first group the populations by cluster analysis (Table 5) and we also find the Prim network (Figure 5, in which the interbranch angles are also given). Cluster analysis leads to form 15, in which the initial split is into Eskimo and Korean on the one branch, and Bantu and English on the other, while, on an examination of the angles, the Prim network suggests the same form, but unrooted (3′). The second best cluster analysis leads to form 6.

The "additive-tree" and "minimum-evolution" results for the three possible unrooted forms are given in Table 6, from which it will be seen that form 3′ is, in both instances, again the best. Table 7 gives the dimensions of form 3′ on the two methods, and on the maximum-likelihood method (see below); the other two forms lead to zero (minimum-evolution) or negative (additive-tree) values for the central branch, thus indicating the need

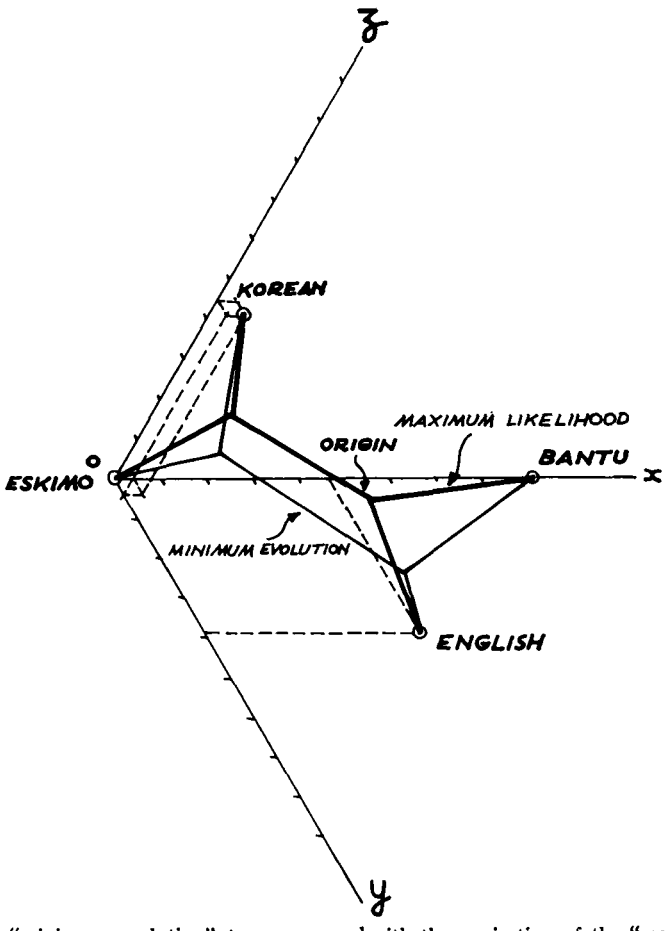


FIG. 6. The "minimum-evolution" tree compared with the projection of the "maximum-likelihood" tree into the "now" space.

for a change of form. The node coordinates for the "minimum-evolution" solution will be found in parentheses in Table 8.

Applying the modified maximum-likelihood method it transpires that only two of the rooted tree forms, numbers 6 and 15, give convergent results, the likelihood of 15 being 3.826 times that of 6. Branches of negative time interval are generated with the remaining thirteen forms. Since form 15 has twice the prior probability of form 6, the final likelihood ratio in its favor is 7.652. Its node coordinates are given in Table 8, and its spatial dimensions in Table 7. Figure 6 shows its projection in the "now" space, together with the "minimum-evolution" tree. In passing, Table 9 gives the node coordinates of the second best "maximum-likelihood" tree, of form 6.

We thus find that all the methods tried lead us to form 15 (or its unrooted equivalent, 3′), and the dimensions of this tree given by the modified method of maximum likelihood represent, we believe, the best estimates of the evolutionary paths on the given data. For the present we are not prepared to quote standard deviations for our estimates because of the complexity introduced by the number of possible forms. The problem is probably best approached by a numerical examination of the consequences of small displacements of the populations in the character space.

TABLE 7. *Dimensions of the best tree (form 3') according to the three methods.*

Branch	Length in gene substitutions		
	"additive tree"	"minimum evolution"	"maximum likelihood"
Eskimo	0.1913	0.1998	0.2451
Korean	0.3005	0.3605	0.2946
Bantu	0.4375	0.5169	0.4895
English	0.1532	0.1270	0.2720
central	0.2752	0.4132	0.3588
Total	*	1.6173	1.6600

"Central" refers to the branch uniting the two splits.

* In comparing these lengths it must be remembered that the "additive tree" cannot be represented in the character space. The comparative shortness of the "central" branch was anticipated in the section on the "method of minimum evolution."

TABLE 8. *Node coordinates for the "maximum-likelihood" and "minimum-evolution" trees of form 15 and 3'.*

Node	*x*	*y*	*z*	Time
top	0.4238	0.1258	0.0972	0.1698
Esk-Kor	0.1229 (0.1221)	0.0482 (0.0747)	0.2065 (0.1394)	0.0694
Ban-Eng	0.4506 (0.4710)	0.1327 (0.2656)	0.0875 (0.0276)	0.1609

("minimum-evolution" coordinates in parentheses)

DISCUSSION

The introduction of automatic computing methods in phylogenetic analysis has the advantage of objectivity, but this is not the greatest advantage. An automatic procedure needs a clearly formulated set of rules, and discussion of the acceptability of the rules requires a knowledge of the aims. Clarification of thought is thus likely to result, and may be more important than the improvement in precision and speed brought about by the automatic procedure. In particular, by introducing a specific model the limitations of the method will be clear for all to see.

The reconstruction of evolutionary trees is a type of inductive inference which is likely to be especially weak. Relevant evidence comes from many sources, and is often conflicting and difficult to weigh; the sheer mass of evidence may be a bar to its interpretation. Under these conditions, an objective method capable, in principle, of assaying the strength of the evidence, of testing the goodness of fit of specific models, and of using various kinds of data, should be considered a step forward.

Ideally, before the use of any such method can be recommended with complete confidence, we would like to be able to state that it gives a satisfactory answer in well-known test cases. The most hopeful sources of data at the moment are gene frequency data in cattle and man, and nucleic acid hybridization and protein structure analyses. In this and previous publications we have used gene frequency data in man both on account of its suitability and its availability. Unfortunately, however, there is not much guidance from other sources as to the recent evolutionary history of man, and the material is thus not very suitable for testing our methods. But we have been encouraged by the reasonableness of results obtained on very modest amounts of data, as, for example, the form of the tree found for fifteen human populations using the same type of data as we have used in the present paper (Edwards and Cavalli-Sforza, 1964; Cavalli-Sforza, Barrai, and Edwards, 1964). In plotting the tree on a map of the world, the actual branch lengths were unavoidably, and admittedly, distorted, and only the form was undisturbed. It is therefore clear that a criticism levelled against this map—that the Maoris seem to "stem phylogenetically from the natives of Alaska" (Simpson, 1965)—is not valid. The branch leading to the Maoris could equally well have been drawn in many other ways without disturbing the basic form, and it should also be remembered that other forms with somewhat similar branching sequences were not much less likely than the one shown on the map: forms, as well as dimensions, are subject to statistical error. But given that

TABLE 9. *Node coordinates for the second best "maximum-likelihood" tree, of form 6.*

Node	x	y	z	Time
Top (Ban)	0.4325	0.0971	0.1040	0.2117
Eng	0.2940	0.1251	0.1339	0.1508
Esk-Kor	0.2620	0.1130	0.1467	0.1415

only a minute fraction of the available data was used, and that no geographical information was incorporated in the estimation procedure, the result seems encouraging.

Among the limitations of our model that have been considered in previous sections, inability to handle hybridization, convergence and parallelism (that is, the similar selective response to similar environmental stimuli in different populations) needs particular consideration. While hybridization gives rise to loops in the tree, which might possibly be detected, convergence and parallelism cause a breakdown of the assumption that evolution proceeds independently on each branch of the tree. This assumption is basic to our model (excepting the case of directional selection constant in space: see Figure 2), and probably basic to any tractable model for evolutionary divergence. In the absence of significant loops, the breakdown of this assumption is likely to be the major cause of poor fits.

However, it seems unlikely that this breakdown can occur in such a way as to cause extensive similarity at the genotypic level between organisms that have diverged greatly in the past and have since evolved in a common environment. Some convergence may be detected at a few loci, in which case a choice of loci might have to be exercised. Unfortunately, so little is known about the selective pattern of most genes that such a choice will be difficult, if not dangerous, most of the time. In man, the same or similar environments, like malaria, can bring about different selective responses, which depend on the genetic background of the exposed population, on its history, and on other environmental factors. In the case of malaria quite a variety of genetic adaptations have been observed, thalassemia, G6PD, and hemoglobins S and E being the best-known examples.

To sum up, we cannot do better than repeat an earlier warning, that "prolonged periods of selection peculiar to individual populations will not be detectable without data from the past [or other information about the selective situation], and no method of phylogenetic analysis can alter the fact than any observed diversity can be explained by any evolutionary tree provided we are willing to postulate the necessary selection." But there is reason to believe that, where enough different genes are considered, the effects of truly convergent genes will be swamped by the larger number of genes behaving independently in different populations.

In the example we have used, the best estimate for the time of the first split was 0.17 units of $2t\sigma^2$ ago, where t is the number of generations and σ^2 the variance per generation. Putting $\sigma^2 = \frac{1}{8}N$, the variance due solely to drift in a population of effective size N, $t = 0.68\ N$ generations. Such estimates should serve not only to increase our knowledge of evolutionary history, but also to help us to understand the evolutionary process. For example, if independent evidence corroborates an estimate of the time taken the model is to some extent vindicated, particularly if it can be shown that the assumed population sizes are sufficient to have maintained the observed numbers of alleles.

SUMMARY

An attempt has been made to establish a procedure for estimating the course taken by evolution. The model used is that of a branching random walk, which is strictly valid only when the causes of divergence between populations are random genetic drift and variable selection. With suitable transformations of the data, evolution can then be considered as a branching Brownian-motion process. To keep the model as simple as possible it was supposed

that no population becomes extinct and that each population splits, at a random time, into two daughter populations each identical to its parent. The problem was to estimate the form and dimensions of the most probable tree uniting the presently living populations. The ideal method of estimation, maximum likelihood, proved difficult and had to be replaced in part by alternative procedures. In addition to describing the available procedures in detail, a simple example is worked out fully, and the logical content and limitations of the methods are considered in depth.

ACKNOWLEDGMENTS

This work has been supported by grants from the U.S. Atomic Energy Commission and by EURATOM-CNR-CNEN Contract No. 012-61-12 BIAI.

We are particularly indebted to Dr. Laura Zonta for her painstaking cooperation in the heavy numerical work involved in the development of our methods, and to the late Professor Vittorio Galafasi and Mr. E. F. Harding for illuminating discussions on tree forms.

The final draft of this paper was prepared after A.W.F.E. had taken up an appointment in the University of Aberdeen.

LITERATURE CITED

CAMIN, J. H., AND R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution **19**: 311–326.

CAVALLI-SFORZA, L. L. 1966. Population structure and human evolution. Proc. Roy. Soc., London, B. **164**: 362–379.

CAVALLI-SFORZA, L. L., AND F. CONTERIO. 1960. Analisi della fluttuazione di frequenze geniche nella popolazione della Val Parma. Atti Assoc. Genet. Ital. **5**: 333–344.

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1964. Analysis of human evolution. Proc. XI Int. Congr. Genet. **3**: 923–933.

——, AND ——. 1966. Estimation procedures for evolutionary branching processes. Bull. Inst. Int. Statist., 35e Session. *In press.*

CAVALLI-SFORZA, L. L., I. BARRAI, AND A. W. F. EDWARDS. 1964. Analysis of human evolution under random genetic drift. Cold Spring Harbor Symp. Quant. Biol. **29**: 9–20.

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

——. 1859. On the analytical forms called trees. Phil. Mag. **18**: 374–378.

CHOWN, B., AND M. LEWIS. 1959. The blood group genes of the Copper Eskimo. Amer. J. Phys. Anthropol. **17**: 13–18.

COURANT, R., AND H. ROBBINS. 1961. What is mathematics? Oxford Univ. Press, Oxford.

EDWARDS, A. W. F. 1967. The shortest network uniting a set of points, additional nodes being allowed. *In preparation.*

EDWARDS, A. W. F., AND L. L. CAVALLI-SFORZA. 1963*a*. The reconstruction of evolution. Abstract in: Ann. Hum. Genet., London **27**: 105 and Heredity **18**: 553.

——, AND ——. 1963*b*. A method for cluster analysis. Preprints of V Internat. Biometric Conf. Abstract in: Biometrics **20**: 383 (1964).

——, AND ——. 1964. Reconstruction of evolutionary trees. Systematics Assoc. Publ. No. 6: Phenetic and Phylogenetic Classification: 67–76.

——, AND ——. 1965. A method for cluster analysis. Biometrics **21**: 362–375.

FISHER, R. A. 1958. Statistical methods for research workers. Thirteenth ed. Edinburgh: Oliver and Boyd.

HARDING, E. F. 1967. The probabilities of root-trees generated by random bifurcation. *In preparation.*

KENDALL, M. G., AND A. STUART. 1961. The advanced theory of statistics, 2, Ch. 19. Griffin, London.

KIMURA, M. 1954. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. Genetics **39**: 280–295.

——. 1955. Random genetic drift in multi-allelic locus. Evolution **9**: 419–435.

LEWIS, M., H. KAITA, AND B. CHOWN. 1957. The blood groups of a Japanese population. Amer. J. Human Genet. **9**: 274–283.

POLYA, G. 1937. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen, und chemische Verbindungen. Acta Math. **68**: 145–253.

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

RACE, R. R., AND R. SANGER. 1959. Blood groups in man. Third ed. Blackwell Scientific Publications, Oxford.

RAO, C. R. 1952. Advanced statistical methods in biometric research. Wiley, New York.

SIMPSON, G. G. 1965. Current issues in taxonomic theory (Book Review). Science **148**: 1078.

SOKAL, R. R., AND P. H. A. SNEATH. 1963. Principles of numerical taxonomy. Freeman, San Francisco.

WARD, J. H. 1963. Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. **58**: 236–244.

WON, C. D., H. S. SHIN, S. W. KIM, J. SWANSON, AND G. A. MATSON. 1960. Distribution of hereditary blood factors among Koreans residing in Seoul, Korea. Amer. J. Phys. Anthropol. **18**: 115–124.

YULE, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. Roy. Soc. London B, **213**: 21–87.

ZOUTENDYK, A., A. C. KOPEĆ, AND A. E. MOURANT. 1955. The blood groups of the Hottentots. Amer. J. Phys. Anthropol. **13**: 691–697.

ZUCKERKANDL, E. 1965. The evolution of hemoglobin. Sci. Amer. **212**: 110–118.