Evolution, 35(6), 1981, pp. 1229–1242

EVOLUTIONARY TREES FROM GENE FREQUENCIES AND QUANTITATIVE CHARACTERS: FINDING MAXIMUM LIKELIHOOD ESTIMATES

JOSEPH FELSENSTEIN
Department of Genetics, University of Washington, Seattle, Washington 98195

Received July 23, 1980. Revised February 28, 1981

A small but complex literature on the estimation of evolutionary trees from quantitative characters (including gene frequencies) has existed for over 15 years (Edwards and Cavalli-Sforza, 1964; Cavalli-Sforza and Edwards, 1967, 1970; Edwards, 1970; Kidd and Sgaramella-Zonta, 1971; Thompson, 1973; Felsenstein, 1973*a*; Thompson, 1975; Cavalli-Sforza and Piazza, 1975; Astolfi et al., 1978). In general its methods are little-known and even less used by those in possession of relevant gene frequencies or quantitative character data. Though this results in part from the complexity of the mathematics in these papers and in part from their concentration in human genetics journals, a major block to the use of these statistical methods has been the difficulty of the computations.

Thompson (1975) has produced an iterative computer program which is probably the most efficient method of finding maximum likelihood evolutionary trees. Thompson's method is the strict application of maximum likelihood estimation in a situation in which each character added to the data brings with it one new parameter to be estimated. It is easily demonstrated that in this case, these "nuisance parameters" cause the estimation procedure to fail to be consistent: that is, the estimate will not converge to the true tree as more and more characters are added. My own procedure (Felsenstein, 1973*a*) makes a restricted maximum likelihood (REML) estimate, but eliminates the presence of the nuisance parameters and thereby makes a consistent estimate of the evolutionary tree. Some of the differences between these approaches will be briefly dealt with later in this paper.

This paper introduces an iterative REML method which makes rapid computation of the REML estimate of the evolutionary tree feasible. It may also serve as a review of the basic logic of these estimates and tests for those readers unfamiliar with the existing human genetics-oriented literature.

THE EVOLUTIONARY MODEL

The primary assumption of all papers in this literature, starting with Edwards and Cavalli-Sforza (1964), is that each character evolves independently according to a Brownian motion process, the mean phenotype in a population undergoing a random diffusion on an infinite linear scale. Figure 1 shows a simulation of evolution along a branching tree of lineages according to this process. There are two sorts of biological processes which could cause a character to evolve in this way.

Random genetic drift will be well-approximated by Brownian motion, except that the rate of diffusion will differ in different parts of the gene-frequency scale. With two alleles, the variance of the change in gene frequency each generation is $p(1 - p)/2N_e$, where p is the frequency of one allele and N_e is the effective population size. As discussed by Thompson (1975), the transformation $\sin^{-1}\sqrt{p}$ effectively removes this inhomogeneity of variance in a way which is satisfactory at all but extreme gene frequencies. The resulting variate accumulates a variance of approximately $1/(8N_e)$ per generation.

One other mechanism which could cause an approximate Brownian motion is random variation of selection coefficients. With two alleles, the variance of

FIG. 1. Schematic view of the Brownian motion model of evolution used in this paper.

the change in gene frequency in one generation is $V_s p^2(1 - p)^2$, where V_s is the variance of the selection coefficient s. The logit transformation $\ln[p/(1 - p)]$ will remove the dependence of this variance on gene frequency.

It is assumed that the variance of the Brownian motion of each character is the same. This is approximately true for genetic drift, especially after transformation to the arc-sine scale, but it could only be true for varying selection if V_s were the same at all loci. This is sufficiently implausible that we cannot invoke variation of selection coefficients as a rationale for Brownian motion.

Removing Correlations

If there are multiple alleles, the individual allele frequencies will not drift independently. The multiple-allele analog of the arc-sine transformation, due originally to Bhattacharyya (1946), has been developed by Cavalli-Sforza and Edwards (1967) into their genetic distance measure. This computes the geometric distance between two populations in a space in which the k alleles trace out a $(k - 1)$-dimensional Brownian motion, the coordinates in the new space being nearly independent in their motions. As we shall see that such distances are sufficient to allow estimation of the tree, this measure is a reasonable solution of the problem posed by multiple alleles.

When we do not observe allele frequencies, but instead the values of a number of quantitative characters, we cannot expect either that these characters drift independently or that they have equal drift variances. Fortunately the covariances of characters in the drift process are expected to be a simple multiple of the within-population additive genetic covariances of these characters. If the additive genetic covariance of characters i and j is σ_{ij}, the covariance of their changes by genetic drift will be approximately σ_{ij}/N_e per generation. If we make a linear transformation of the characters to a set of new variables which have unit additive genetic variances and no additive genetic covariance, these new variables will drift independently and at an equal expected rate.

The problems caused by the finiteness of population samples in the estimation of these covariances have not yet been seriously investigated, nor have methods of estimating additive genetic covariances in the types of data likely to be encountered in practice. In particular, environmental factors which can differ between populations or can be transmitted from one generation to the next will considerably complicate this picture. In practice, the requirement to know additive genetic covariances between characters will restrict application of the present methods to gene frequency data. Additive genetic covariances can be estimated for quantitative characters from between-population information. The present framework serves as the starting point for that method, which will be described elsewhere.

The Distribution of the Data

In the Brownian motion model, the net change after t units of time is normally distributed with zero mean and variance ct, with the constant c being the same for all characters. The change in two lineages

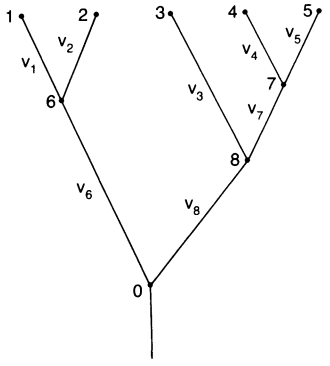

FIG. 2. A five-population tree, demonstrating the notation used in this paper.

after speciation is assumed independent of each other, and the change of a character during two successive intervals of time is also independent. Figure 2 shows a five-population evolutionary tree, with the interior nodes (the branch points) numbered as well. Next to each segment of the tree is the amount of variance (v_1 through v_8) which is expected to accumulate during evolution along that part of the tree. Suppose that the values of the mean phenotype at points 1 through 0 are given by $x_1, \ldots, x_8, x_0$. We can determine the joint distribution of the observable values $x_1, \ldots, x_5$ given x_0 by the following argument.

The value of x_3 is

$$x_3 = (x_3 - x_8) + (x_8 - x_0) + x_0 \quad (1)$$

and the value of x_4 is

$$x_4 = (x_4 - x_7) + (x_7 - x_8) + (x_8 - x_0) + x_0. \quad (2)$$

The expression for x_3 is the sum of three terms, each of the first two normally distributed, and by assumption independent of each other. Their expectations are zero and their theoretical variances v_3 and v_8. So x_3 is normally distributed with expectation x_0 and variance $v_3 + v_8$. A similar argument establishes that x_4 is normally distributed with expectation x_0 and variance $v_4 + v_7 + v_8$. The two quantities x_3 and x_4 are jointly bivariate normally distributed, with a covariance due to their common term $(x_8 - x_0)$. Their covariance will be the variance of this common term, v_8.

The general pattern is that $x_1, \ldots, x_5$ are multivariate normally distributed, with expectations all being x_0 and with their covariances being the variances which accumulate during the parts of the tree shared by the ancestry of the two particular populations. Thus the covariance of x_1 and x_2 is v_6, that of x_1 and x_3 is zero, that of x_4 and x_5 is $v_7 + v_8$, and so on. The variance of each value is the sum of the v's along its ancestral lineage leading from point 0. This multivariate normality was discovered by Gomberg (1966).

Note that the lengths of the tree segments in time affects this distribution only through the v_i. Each of the v_i is of the form $c(t_i - t_j)/N_e$, where $t_i - t_j$ is the length of segment i in time, c is the within-population additive genetic variance, and N_e is the effective population size during tree segment i. Unless the effective population sizes are known, they and the times are inextricably confounded. If (say) v_7 appears to be small, this may result from a small span of evolutionary time, or a large effective population size, or both. We are measuring opportunity for random change, not time itself.

If population i is represented by a sample of n_i organisms, the resulting sampling error in estimating the population mean is exactly the same as one generation of genetic drift at an effective population size of n_i. We can take account of this by adding cT/n_i to the expected variance of population i, where T is the length of a generation. This does not properly correct for the effect of small sample size on any estimates we have made of additive genetic covariance.

In this treatment we take the form of the evolutionary tree as the starting point, without attempting to make a probabilistic model of the branching and extinction of lineages. Thompson (1975) has given statistical reasons for not including such a model in the analysis (where it would serve to provide prior information on the

quantities being estimated). There are biological reasons as well: we would have to incorporate into such a model not only probabilities of birth of new lineages and extinction of old ones, but also a process corresponding to the random or arbitrary selection of lineages to be studied. This last process seems hard to model.

We will be concerned here with maximum likelihood methods, because these make fully efficient use of the data and allow likelihood-ratio tests of hypotheses. The least-squares approach of Cavalli-Sforza and Edwards (1965; see explanation in Kidd and Sgaramella-Zonta, 1971) will have some desirable statistical properties such as consistency. So will the reconstruction method of Malyutov et al. (1971), which lacks a general criterion for choosing among tree topologies. The most common method of treating gene-frequency data on evolutionary trees has been to apply the average-linkage clustering method to a matrix of genetic distances. This too will make a consistent estimate of the tree if the genetic distance is expected to rise linearly with time and if the rates of evolution are equal in each lineage. Unequal rates of evolution (as when effective population sizes differ among lineages) can cause incorrect estimates of the topology even with an infinite amount of data, as noted by Colless (1970). The method of estimation and testing recently introduced by Cavalli-Sforza and Piazza (1975) uses covariances between populations and a maximum likelihood method, but does not use maximum likelihood for estimating the means around which the covariances are computed. This problem seems connected to the discrepancy between their approach and the maximum likelihood method of Thompson (1975).

The Case of Two Populations

Maximum likelihood solution.—All of the statistical and computational properties which we will need are visible in a simple analysis of the two- and three-population cases. In the two-population case, where the evolutionary tree is as given in Figure 3, the likelihood function is a product of terms, one for each character, each a product of two independent normal distributions, one with mean x_0 and variance v_1, the other with mean x_0 and variance v_2. If we have p characters and let x_{ij} be the value of character j in population i,

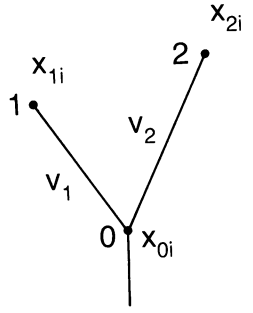

FIG. 3. The two-population tree used for exact solution of the likelihood equations.

$$L = \prod_{j=1}^{p} \frac{1}{\sqrt{(2\pi v_1)}} \exp\left[-\frac{(x_{1j} - x_{0j})^2}{2v_1}\right] \cdot \frac{1}{\sqrt{(2\pi v_2)}} \exp\left[-\frac{(x_{2j} - x_{0j})^2}{2v_2}\right] \quad (3)$$

which means that the log likelihood is

$$\ln L = \text{constant} - \frac{p}{2}\ln v_1 - \frac{p}{2}\ln v_2 - \frac{1}{2v_1}\sum_j (x_{1j} - x_{0j})^2 - \frac{1}{2v_2}\sum_j (x_{2j} - x_{0j})^2. \quad (4)$$

If we differentiate this expression with respect to each of its parameters $x_{01}, \ldots, x_{0p}$, v_1, and v_2, and equate the resulting expressions to zero, we find that their maximum likelihood estimates are the solutions to

$$\hat{x}_{0j} = \frac{(1/\hat{v}_1)x_{1j} + (1/\hat{v}_2)x_{2j}}{1/\hat{v}_1 + 1/\hat{v}_2}, \quad (5)$$

and

$$\hat{v}_1 = \textstyle\sum_j (x_{1j} - \hat{v}_{0j})^2/p, \quad (6a)$$

$$\hat{v}_2 = \textstyle\sum_j (x_{2j} - \hat{x}_{0j})^2/p. \quad (6b)$$

These are quite straightforward expressions. The estimate of x_{0j} is the weighted

average of the values x_{1j} and x_{2j}, and v_1 is estimated by the mean square of the x_{1j} around their estimated means x_{0j}. However, equations (5) and (6) do not determine $\hat{x}_{0j}$, $\hat{v}_1$, and $\hat{v}_2$ uniquely. We find after some algebra that the maximum likelihood is obtained by letting one of the $\hat{v}_i$ (say $\hat{v}_1$) approach zero, in which case

$$\hat{x}_{0j} = x_{1j} \tag{7a}$$

and

$$\hat{v}_2 = \sum_j (x_{1j} - x_{2j})^2/p. \tag{7b}$$

The problem with this solution is that as $\hat{v}_1$ becomes zero, the likelihood becomes infinite! We then have no basis for comparing likelihoods of the two solutions in which $\hat{v}_1 = 0$ or $\hat{v}_2 = 0$.

A singularity of the likelihood surface similar to this was observed by Edwards and Cavalli-Sforza (1964) and formed the basis of their abandonment of a likelihood approach. The matter has been discussed by Thompson (1975), who concluded that their problem arose from failing to distinguish between parameters and random variables. That conclusion seems not to apply to the present example.

The problem could be avoided by constraining v_1 and v_2. When we constrain v_1 and v_2 to be equal, the singularity in the likelihood surface no longer exists, and we find that the estimates are simply

$$\hat{x}_{0i} = \tfrac{1}{2}x_{1j} + \tfrac{1}{2}x_{2j} \tag{8a}$$

and

$$\hat{v}_1 = \hat{v}_2 = \sum_j (x_{1j} - x_{2j})^2/(4p). \tag{8b}$$

Thompson's (1975) maximum likelihood approach leads to this result. Other types of constraints can also be used.

Inconsistency of the estimate.—The difficulty with the solution (8), aside from the arbitrariness of the constraint, is that it leads to a grossly inaccurate result. When the true values of v_1 and v_2 are both v, $(x_{1j} - x_{2j})^2$ has an expectation of $2v$, so that the estimates $\hat{v}_1$ and $\hat{v}_2$ each have expectations of $v/2$. This bias does not disappear as we observe more and more characters. In fact, the estimates converge to $v/2$ as more data are accumulated.

One normally expects maximum likelihood estimates to be better-behaved, in particular to converge with certainty to the correct value of the parameter as more data are accumulated. The reason for this lack of statistical consistency is the presence of the "nuisance parameters" x_{0j}. As we add more characters to the analysis we also add more parameters. The ratio of observations to parameters never rises above a fixed value (in this case 2). In fact, this very case (though without its taxonomic interpretation) is a classical counterexample to the use of likelihood (see discussion in Kendall and Stuart, 1973, p. 63). It is intriguing that a similar inconsistency arises when maximum likelihood is used on discrete-character data (Felsenstein, 1973*b*, 1978*b*, 1979).

Restricted maximum likelihood.—We can escape from this problem by noting that it seems on intuitive grounds that all the information about the v_i is contained in the differences $x_{1j} - x_{2j}$, and the actual positions of x_{1j} and x_{2j} on the scale of character i serve only to help us estimate the nuisance parameter x_{0j}. Suppose that we assume that we have observed only the differences between populations, and not their absolute positions on each character's scale. In each character, $x_{1j} - x_{2j}$ is drawn from a normal distribution with mean zero and variance $v_1 + v_2$. The likelihood function of this restricted set of data is

$$L = \prod_{j=1}^{p} \frac{1}{\sqrt{[2\pi(v_1 + v_2)]}} \cdot \exp\left[-\frac{(x_{1j} - x_{2j})^2}{2(v_1 + v_2)}\right]. \tag{9}$$

The v_i enter into (9) only through their sum $v_1 + v_2$. There are no nuisance parameters. Taking logarithms and differentiating, the maximum likelihood estimates are

$$\hat{v}_1 + \hat{v}_2 = \sum_{j=1}^{p} (x_{1j} - x_{2j})^2/(2p). \tag{10}$$

There are no singularities in the likelihood

surface. This estimate of $v_1 + v_2$ is unbiased, and guaranteed to converge to the true value as we accumulate more and more characters.

This restricted maximum likelihood (REML) seems to deliver an acceptable estimate with none of the pathologies associated with the full maximum likelihood (ML) approach. Furthermore we are guaranteed many properties usually associated with maximum likelihood estimates, for the REML estimate of the evolutionary tree *is* a maximum likelihood estimate, though one obtained after the data have been passed through a filter.

Restricted maximum likelihood has been used often in statistics. In the estimation of variance components it was introduced by Patterson and Thompson (1971). Harville (1977) gives a good comprehensive review of its use in the analysis of variance, though the present cases do not fall within his framework. The present REML approach was introduced in my earlier paper (1973*a*) on this subject, and also in remarks by Gomberg (unpubl.).

Three Populations

Most of the remaining complications can be seen in the three-population case. Suppose that we try to estimate the parameters v_1, v_2, v_3, and v_4 of the tree shown in Figure 4. The three observed values for each character (x_{1j}, x_{2j}, and x_{3j}) are expected to be multivariate normally distributed with means (x_{0j}, x_{0j}, x_{0j}) and covariance matrix.

$$\begin{bmatrix} v_1 + v_4 & v_4 & 0 \\ v_4 & v_2 + v_4 & 0 \\ 0 & 0 & v_3 \end{bmatrix}.$$

The calculation of likelihoods is considerably simplified (and the issues separating different methods clarified) if we consider differences between populations. The difference $x_{1j} - x_{2j}$ has expectation zero and variance

$$(v_1 + v_4) + (v_2 + v_4) - 2v_4 = v_1 + v_2. \qquad (11)$$

The value x_{3j} is independent of $x_{1j} - x_{2j}$,

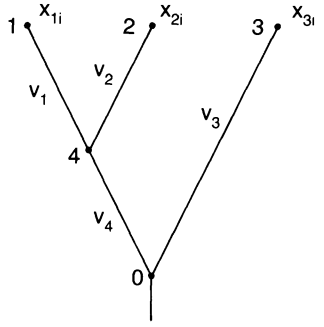

FIG. 4. The three-population tree used to demonstrate the pulley principle and the pruning algorithm.

and if we make up the linear combination

$$x_{4j}' = \frac{(1/v_1)x_{1j} + (1/v_2)x_{2j}}{1/v_1 + 1/v_2}, \qquad (12)$$

it will have zero covariance with $x_{1j} - x_{2j}$ and also with x_{3j}. Thus the three values ($x_{1j} - x_{2j}$, x_{4j}', x_{3j}) are independently normally distributed. Their expectations are (0, x_{0j}, x_{0j}).

To make an REML estimate of the tree we note that we are not interested in the values x_{4j}' and x_{3j}, but only in their difference $x_{4j}' - x_{3j}$. This difference is normally distributed with mean zero. Its variance is the sum of v_3 and the variance of x_{4j}', and the latter has variance which may be calculated to be

$$v_4' = v_4 + 1/(1/v_1 + 1/v_2). \qquad (13)$$

In fact, $x_{4j}' - x_{3j}$ is only dependent on the differences between the original populations, as it is the weighted average of $x_{1j} - x_{3j}$ and $x_{2j} - x_{3j}$, the weights being $1/v_1$ and $1/v_2$.

Pruning the tree.—This process of removing nodes 1 and 2 from the tree and replacing them with the fictional values x_{4j}' was introduced in my earlier paper (Felsenstein, 1973*a*) and called "pruning." If we are making an REML estimate of the tree, the likelihood of the tree in Figure 4 is exactly the same as the product of the likelihoods of the two trees in Figure 5. By likelihoods we hereafter mean the restricted likelihood based on differences, unless otherwise specified.

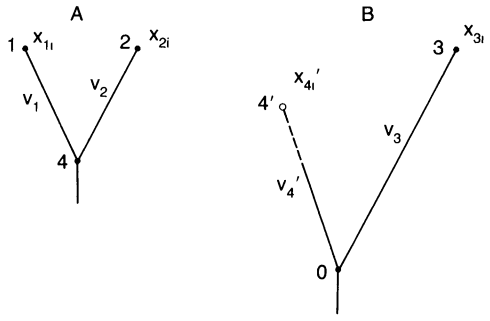

FIG. 5. The pruning algorithm. These two trees, taken together, have the same restricted likelihood as the tree of Figure 3. x_{4i}' and v_4' are defined in the text.

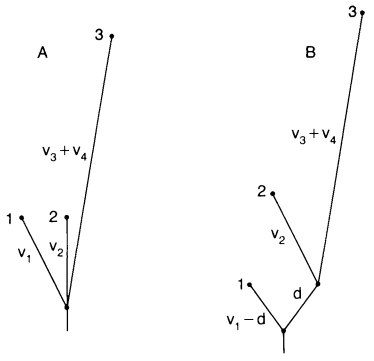

FIG. 6. The pulley principle. Both trees (A) and (B) have the same restricted likelihood as the tree in Fig. 3.

The pruning algorithm is closely related to the "peeling" algorithm introduced by Elston and Stewart (1971) for the computation of likelihoods from pedigrees in human statistical genetics. There the emphasis is usually not on inferring pedigree form but on fitting models of inheritance to known pedigrees.

Pruning can be used for the rapid computation of restricted likelihoods, and the general algorithm for this was given in my earlier paper (Felsenstein, 1973*a*). The algorithm proceeds by removing one pair of populations at a time from the tree, leaving behind a single fictional population each time. After $p - 1$ such removals, we have $p - 1$ independent trees. Their joint likelihood, which is an easily-evaluated product, is equal to the likelihood of the original tree. The differences and weighted averages are constructed in a way which ensures that no problems arise from the Jacobians of these transformations.

The pulley principle.—Note that the likelihood of the tree in Figure 4 will depend on v_3 and v_4 only through their sum $v_3 + v_4$. This can be seen by noting that these two quantities appear in Figure 5 only in the right-hand tree, whose likelihood is

$$L = \prod_{j=1}^{p} \frac{1}{\sqrt{[2\pi(v_4' + v_3)]}} \cdot \exp\left[-\frac{(x_{4j}' - x_{3j})^2}{2(v_4' + v_3)}\right], \qquad (14)$$

and that by (13) this contains v_3 and v_4 only through their sum.

Thus, as far as REML estimation is concerned, the bottom fork of the tree acts as a sort of pulley. We can add an amount d to v_4, and subtract the same amount from v_3, and in doing so we leave the likelihood unchanged. The tree in Figure 6A will have the same likelihood as that in Figure 4. Furthermore, once a three way fork is reached, as in Figure 6A, the pulley can be rolled onto any branch. Figure 6B shows another tree which will have the same likelihood as the trees of Figures 4 and 6A.

Thus the pulley principle identifies a large class of trees which will all have equal likelihoods. REML estimation can only give us information about which of these equivalence classes may contain the correct tree—it cannot carry the matter any further without other sources of information. Each equivalence class of trees consists of all the information about the trees except for the position of the root and the first fork. Thus what we are estimating is an unrooted tree connecting the populations, and the root may be placed anywhere in this unrooted tree.

Exact solution of the three-population case.—The pruning process and the pulley principle can be combined to obtain an expression for the likelihood of a three-population tree. This could also be ob-

tained by direct consideration of the likelihood function, though with greater difficulty. It is assumed that the pulley principle has been applied to a three-population tree to bring it into the form in Figure 7. Appendix 1 then demonstrates that the likelihood can be written as

$$L = \frac{1}{(2\pi)^p[(v_1v_2 + v_1v_3 + v_2v_3)]^{p/2}} \cdot \exp\left[-\frac{v_1D_{23}^2 + v_2D_{13}^2 + v_3D_{12}^2}{2(v_1v_2 + v_1v_3 + v_2v_3)}\right] \quad (15a)$$

where

$$D_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \quad (15b)$$

is the squared geometric distance between populations i and j in the coordinate system defined by the independent characters.

Appendix 2 demonstrates the derivation of the maximum likelihood estimates $\hat{v}_1$, $\hat{v}_2$, and $\hat{v}_3$ which turn out to be

$$\hat{v}_1 = (D_{12}^2 + D_{13}^2 - D_{23}^2)/(2p) \quad (16a)$$
$$\hat{v}_2 = (D_{23}^2 + D_{12}^2 - D_{13}^2)/(2p) \quad (16b)$$
$$\hat{v}_3 = (D_{13}^2 + D_{23}^2 - D_{12}^2)/(2p) \quad (16c)$$

and which can also be written as

$$\hat{v}_1 = \sum_j (x_{1j} - x_{2j})(x_{1j} - x_{3j})/p \quad (17a)$$

$$\hat{v}_2 = \sum_j (x_{2j} - x_{1j})(x_{2j} - x_{3j})/p \quad (17b)$$

$$\hat{v}_3 = \sum_j (x_{3j} - x_{1j})(x_{3j} - x_{2j})/p. \quad (17c)$$

If the points 1, 2, and 3 form an obtuse triangle, one of the $\hat{v}_i$ may be negative. Negative values make no biological sense. We are interested in maximizing the likelihood under the constraint that each $v_i \geq 0$. If (say) $\hat{v}_1$ would be negative from (17a), it can be shown that the maximum likelihood is obtained by setting

$$\hat{v}_1 = 0 \quad (18a)$$
$$\hat{v}_2 = D_{12}^2/p = \sum_j (x_{1j} - x_{2j})^2/p \quad (18b)$$

and

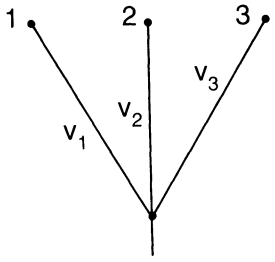

FIG. 7. The tree used for discussion of the exact REML solution of a three-population tree.

$$\hat{v}_3 = D_{13}^2/p = \sum_j (x_{1j} - x_{3j})^2/p. \quad (18c)$$

Such cases are not expected in data generated by evolution on a tree, but might result from sampling error or if either the model of Brownian motion or a tree-like genealogy is inappropriate.

MANY POPULATIONS

The search strategy.—With the pruning algorithm and the pulley principle in hand, we are now ready to construct a method for finding the maximum likelihood tree when there are more than three populations. We will use five populations as our example, but the procedure presented here works generally for any number of populations. The core of the method takes a tree of given topology and makes changes in the lengths of its segments, changes guaranteed to increase the likelihood of the tree. This procedure is iterated, making many changes of segment lengths in different parts of the tree until there is no further change in these lengths. It is then assumed that we have found the combination of lengths which gives the highest likelihood within that given tree topology.

This procedure, described below, is a reasonable solution to the problem of maximizing the likelihood within a given tree topology. But there remains the problem of searching among different tree topologies. It would in principle be possible to try all possible tree topologies, but the number of these is enormous (Felsenstein, 1978*a*). The best that can be done in prac-

tice seems to be to try small changes in the topology until we find one which cannot be improved on. This is not guaranteed to be the true maximum of the likelihood, but it is at least a local maximum.

To succeed with such a strategy, it is necessary to start with a reasonably good initial tree. One procedure which I have found works fairly well is to start with a tree containing only the first three populations, and then add one population at a time to the tree. When the k-th population is to be added, there will be $2k - 5$ internal segments from which it could arise. Each of these is tried, and the one resulting in the highest likelihood accepted. A round of local rearrangement of the topology is then carried out, searching for improvements in the likelihood. When the local maximum for the tree with populations 1 through k is found, population $k + 1$ is ready to be added by the same procedure. Note that this procedure is guaranteed to put population n in the correct place if populations $1, \ldots, n - 1$ are themselves placed correctly. In practice this search procedure seems quite effective. It requires evaluation of at least $2n^2 - 7n + 5$ topologies. Each of these requires iteration of segment lengths. Note that the procedure is sensitive to the order in which the populations are added to the tree. This apparent disadvantage is actually an advantage, as it permits us to search for alternative solutions by trying different orders of populations in the input data. In practice, it is wise to make a number of runs with different orders of populations.

Iterating the lengths.—Within a particular topology, there is a method which will successively alter the lengths of segments, each alteration being guaranteed to increase the likelihood. The tree in Figure 8A will illustrate the algorithm. Suppose that we want to find improved values of the lengths of the three segments which connect to node 8, v_6, v_7, and v_3. We can remove tips 1 and 2, and also tips 4 and 5, from the tree by applying the pruning method twice. The result will be as shown in Figure 8B. Nodes 6 and 7 now have

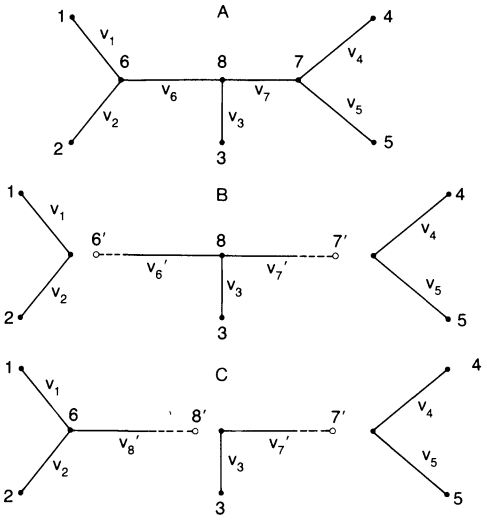

FIG. 8. Unrooted tree (A) has the same likelihood as the three trees in (B) and as the three trees in (C), which have been produced using the pruning algorithm and the pulley principle.

fictional phenotypic values, computed by using a formula exactly analogous to equation (12) above, differing only in the numbering of the nodes. The lengths of segments 6 and 7 have also been altered from v_6 to v_6' and from v_7 to v_7' by application of the pruning method, using equations like equation (13).

The likelihood of the tree in Figure 8A is the product of three terms, one for each of the three trees in Figure 8B. Note that the likelihood depends on v_3, v_6, and v_7 only through the lengths (v_3, v_6', and v_7') of the segments of one three-population tree. If we adjust v_3, v_6', and v_7' so as to be the maximum likelihood solution for this three-population tree, then the likelihood of the five-population tree must be increased, since we have increased the likelihood of one of its components and since v_3, v_6, and v_7 do not enter into the likelihoods of the other components.

If v_1, v_2, v_4, and v_5 were already at their maximum likelihood values, then the adjustment of v_3, v_6, and v_7 would result in the maximum likelihood tree for the given topology. If not, we can invoke the pulley principle to shift attention to (say)

node 6 of the five-population tree. Figure 8C shows the three component trees into which the pruning algorithm decomposes the tree if we take node 6 to be its root. Again, the likelihood depends on v_1, v_2, and v_6 only through the lengths v_1, v_2, and v_8'. Using the maximum likelihood solution for this three-population tree we can find improved values of v_1, v_2, and v_6. We can continue moving through the tree, adjusting lengths, as follows:

1. Move to an interior node.
2. Use the pruning algorithm to decompose the tree into two-population trees, plus one three-population tree centered on the node you chose.
3. Find the restricted maximum likelihood solution for this tree, using equations (16) or (17).
4. From the new lengths of these segments of the three-population tree, find the new lengths of the corresponding segments of the full tree.
5. Move to another interior node, and go back to step 2.

Some economies can be realized by moving from a node to an adjacent node in step 5, since the steps of the pruning algorithm are then quite similar, and need not be redone. In Figures 8B and 8C, the procedure for pruning of populations 4 and 5 will be identical, and that part of the calculation need not be repeated.

This process of moving around the tree, finding improved lengths of segments is guaranteed not to get into an endless loop, since the likelihood is continually increasing and thus we can never return to an earlier state. The process will finally converge, and when the likelihood or the segment lengths essentially stop changing, we can consider ourselves to have arrived at a local maximum of the likelihood within the given tree topology.

There is one problem which may be encountered. It is possible that the new value of (say) v_6' is so small that the new value of v_6 must have been negative. This would be the case if, in the tree of Figure 8B, the new value of v_6' is less than $v_1v_2/(v_1 + v_2)$. When this happens, segment 6 has shrunk to zero length. This indicates that it may be worthwhile to examine topological rearrangements of the tree in this vicinity, as we have been unable to find a maximum of the likelihood within the given tree topology.

Unfortunately, we cannot rely on this phenomenon to point out where the tree needs rearranging. Thompson (1975) has shown that with the restricted likelihood approach there may be local maxima of the likelihood within many different topologies, only one of which is the overall maximum. In Thompson's strict maximum likelihood approach, often the only maximum internal to a topology is the global maximum. This is a computational advantage, but it does not eliminate the other disadvantages of that approach pointed out above. When we have found a maximum of the restricted likelihood internal to a given topology, we are also not guaranteed that it is the only maximum within that topology, although cases where it is not seem fairly unusual.

This method of successive improvements of the tree, each improvement involving a single internal node, is strongly reminiscent of the strategy of Thompson's (1973) algorithm for achieving a "minimum evolution" solution, although both the details of Thompson's algorithm and the results obtained differ.

A computer program.—A computer program to find REML evolutionary trees, written by Mark Moehring, is available from the author. The program is written in PASCAL. It is available as part of a package of programs for numerical estimation of evolutionary trees. This package will be supplied on request, written in standard ANSI format on a magnetic tape supplied by the recipient.

An example.—As a numerical example, Table 1 and Figure 9 show the data and resulting REML tree for five human populations. Gene frequencies of one allele each at ten blood group loci or electrophoretic polymorphisms were extracted from Mourant et al. (1976) for five human populations representative of human geographical variation, namely Western Eu-

TABLE 1. *Gene frequencies for five human groups, taken from Mourant et al. (1976).*

Population	Allele									
	A	*M*	*Rh(CDe)*	*Fya+*	*Hp*[1]	*Gc*[1]	*P*[a]	*PGD*[c]	*PGM*$_1$[1]	*PTC(T)*
West European	0.2868	0.5684	0.4422	0.4286	0.3828	0.7285	0.6386	0.0205	0.8055	0.5043
West African	0.1356	0.4840	0.0602	0.0397	0.5977	0.9675	0.9511	0.0600	0.7582	0.6207
Chinese	0.1628	0.5958	0.7298	1.0000	0.3811	0.7986	0.7782	0.0726	0.7482	0.7334
American Indian	0.0114	0.6990	0.3280	0.7421	0.6606	0.8603	0.7924	0.00	0.8086	0.8636
Aboriginal Australian	0.1211	0.2274	0.5821	1.0000	0.2018	0.9000	0.9837	0.0396	0.9097	0.2976

ropeans, West Africans, Chinese, North American Indian, and aboriginal Australian populations. The populations representing these areas were usually (respectively) those of Paris, Ibadan (Yoruba), Hong Kong, the Zuni people, and Alice Springs. Because not all loci were sampled for all populations, it was necessary to substitute values from alternative populations at some loci. Only one allele frequency was used at each locus. The gene frequencies (which are given in Table 1) were transformed by the arc-sine $\sqrt{p}$ transformation. The tree found by the above mentioned computer program is presented in Figure 9. Although this tree is primarily a computational example, the nearly linear arrangement of populations in the resulting tree is striking. If North American Indian populations were assigned a position in central Asia, this pattern would be quite consistent with a linear pattern of gene flow, and would not necessarily reflect a simple branching pattern.

Approximate confidence limits were calculated for the segments of the tree. For the two segments critical to the topology of the tree these supported the reality of this topology: the confidence limits on the length of the segment separating the West African-European group from the others were 0.0048 to 0.1214, and the confidence limits on the length of the segment separating the Chinese-Australian group from the others were 0.0024 to 0.1044. Both exclude zero.

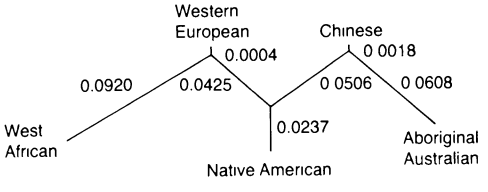

FIG. 9. REML tree for five human populations based on the data in Table 1.

Confidence sets.—One of the great advantages of the maximum likelihood framework is that approximate variances of the quantities estimated can be calculated, and likelihood ratio tests of hypotheses can be carried out. Since the REML estimates are maximum likelihood estimates given an altered set of data, they will allow calculation of asymptotic variances and use of likelihood ratio tests. Of course, these variances and tests assume the validity of the Brownian motion model of evolution, and will not indicate the additional uncertainty caused by our doubts as to the validity of this model.

The variances and covariances of the quantities estimated are computed by inverting a matrix whose elements are $-\partial^2 \ln L/\partial v_i \partial v_j$, the v_i being the quantities estimated. These second derivatives are evaluated at the maximum likelihood tree. Note that there are no quantities corresponding to the topology of the tree: the v_i are the segment lengths within a given topology. The theory which yields variance estimates and likelihood ratio tests is a large-sample asymptotic theory, and a consequence of this is that it treats only variation in the v_i. In effect, this theory assumes that so much data is available that the estimate of the tree must have the correct topology, so that we need only

worry about small variations in the segment lengths.

We can use the covariance matrix of the v_i to construct an approximate joint confidence interval on any set of the v_i. If we are interested in seeing whether an alternative topology can be rejected, this should be roughly the same as asking whether any of the interior segments of the tree can be of zero length. By interior segments, we mean those connecting two interior nodes. When one of these is of zero length, we have a four-way split in the unrooted tree, or a trifurcation if the tree is considered to be rooted. Such a tree is the intermediate between three topological forms. If the joint confidence interval on the lengths of the $n - 3$ interior segments does not allow any of them to be of length zero, alternative topologies can be rejected. This will rarely be possible without enormous amounts of data.

The second derivatives of the log-likelihood can be evaluated numerically by using second-order differences:

$$\frac{\partial^2 \ln L}{\partial v_i \, \partial v_j} \cong \frac{1}{x^2}[F(v_i + x, v_j + x) - F(v_i + x, v_j) - F(v_i, v_j + x) + F(v_i, v_j)] \quad (19)$$

where F is the log-likelihood as a function of v_i and v_j, and x is sufficiently small to allow differences to approximate derivatives.

Likelihood ratio tests.—The likelihood ratio test of a hypothesis H_0 involves finding the overall maximum L of the likelihood, as well as the maximum L_0 under the given hypothesis. The value $-2 \ln (L_0/L)$ will have a Chi-square distribution if the amount of data is sufficiently large, and the hypothesis H_0 can be rejected if this quantity is in the upper tail of the Chi-square distribution. The number of degrees of freedom used in the test is equal to the number of parameters which must be specified to make H_0 true. An important qualification to the test is that the maximum likelihood tree corresponding to the likelihood L be in the interior of a space of possible trees. There are a number of hypotheses of biological interest which may be tested using the likelihood ratio test.

Chief among these is the hypothesis of a constant rate of evolution, which in the present context amounts to the assumption that effective sizes of the populations are all equal, and have been equal in the past. This places a series of constraints on the v_i. The sum of the v_i from each tip down to the bottom node of the tree will be equal under this hypothesis. The iteration scheme presented above does not permit this constraint to be maintained. We do not yet have a computationally effective means of maximizing the likelihood under this restriction.

Summary

The assumptions involved in maximum likelihood estimation of evolutionary trees from quantitative character data have been described. A strict maximum likelihood method applied to the case of two populations encounters singularities in the likelihood surface, and even when restrictions are placed on the parameters to avoid this the resulting estimate converges to the wrong value as more characters are considered. These problems arise because new "nuisance" parameters are introduced every time we add a new character. If the data are assumed to consist only of the differences between population phenotypes, and a maximum likelihood solution based on these transformed data is found, this restricted maximum likelihood (REML) method behaves well. An exact solution is given for the three-population case.

Two computational techniques, the pruning algorithm and the pulley principle, have been described which allow rapid computation of the restricted likelihood. They allow construction of an iterative procedure for finding the maximum of the restricted likelihood within a given tree topology. Combined with an algorithm for searching among similar tree topologies, this allows construction of a

computer program to find the REML estimate of the tree.

ACKNOWLEDGMENTS

I wish to thank Elisabeth Thompson, Larry Mueller, Masatoshi Nei, Ken Kidd and Ryk Ward for many useful comments. The excellent programming talents of Mark Moehring and Jerry Shurman are also much appreciated. This work was supported by test agreement DE-AT06-76EV71005 of contract DE-AM06-76RL02225 between the U.S. Department of Energy and the University of Washington.

LITERATURE CITED

ASTOLFI, P., A. PIAZZA, AND K. K. KIDD. 1978. Testing of evolutionary independence in simulated phylogenetic trees. Syst. Zool. 27:391–400.

BHATTACHARYYA, A. 1946. On a measure of divergence between two multinomial populations. Sankhyā 7:401–406.

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1965. Analysis of human evolution, p. 923–933. *In* S. J. Geerts (ed.), Genetics Today, Proceedings of the 11th International Congress of Genetics, Vol. 3. Pergamon Press, Oxford.

———. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 32:550–570 (also Amer. J. Hum. Genet. 19:233–257).

———. 1970. Estimation procedures for evolutionary branching processes. Bull. Intl. Stat. Inst. 35:803–808.

CAVALLI-SFORZA, L. L., AND A. PIAZZA. 1975. Analysis of evolution: evolutionary rates, independence, and treeness. Theoret. Pop. Biol. 8:127–165.

COLLESS, D. H. 1970. The phenogram as an estimate of phylogeny. Syst. Zool. 19:352–362.

EDWARDS, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. J. Roy. Stat. Soc. B 32:155–174.

EDWARDS, A. W. F., AND L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees, p. 67–76. *In* V. H. Heywood and J. McNeill (eds.), Phenetic and Phylogenetic Classification. Systematics Association Publication No. 6, London.

ELSTON, R. C., AND J. STEWART. 1971. A general model for the analysis of pedigree data. Hum. Hered. 21:523–542.

FELSENSTEIN, J. 1973*a*. Maximum-likelihood estimation of evolutionary trees from continuous characters. Amer. J. Hum. Genet. 25:471–492.

———. 1973*b*. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

———. 1978*a*. The number of evolutionary trees. Syst. Zool. 27:27–33.

———. 1978*b*. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

———. 1979. Alternative methods of phylogenetic inference and their interrelationship. Syst. Zool. 28:49–62.

HARVILLE, D. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Stat. Assoc. 72:320–340.

KENDALL, M. G., AND A. STUART. 1973. The Advanced Theory of Statistics, Vol. 2. Inference and Relationship. Hafner, N.Y.

KIDD, K. K., AND L. A. SGARAMELLA-ZONTA. 1971. Phylogenetic analysis: concepts and methods. Amer. J. Hum. Genet. 23:235–252.

MALYUTOV, M. B., V. P. PASSEKOV, AND Y. G. RYCHKOV. 1972. On the reconstruction of evolutionary trees of human populations resulting from random genetic drift, p. 48–71. *In* J. S. Weiner and J. Huizinga (eds.), The Assessment of Population Affinities in Man. Clarendon Press, Oxford.

MOURANT, A. E., A. C. KOPÉC, AND K. DOMANIEWSKA-SOBCZAK. 1976. The Distribution of Human Blood Groups and Other Polymorphisms, 2nd ed. Oxford Univ. Press, London.

PATTERSON, H. D., AND R. THOMPSON. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554.

THOMPSON, E. A. 1973. The method of minimum evolution. Ann. Human Genet. 36:333–340.

———. 1975. Human Evolutionary Trees. Cambridge Univ. Press, Cambridge.

APPENDIX 1

Restricted Likelihood Function in the Three-Population Case

Application of the pruning algorithm to the tree of Figure 6 discloses that the (restricted) likelihood of that tree is the product of two parts. One reflects the terms $x_{1j} - x_{2j}$, which are $N(0, v_1 + v_2)$, the other values $(v_2x_{1j} + v_1x_{2j})/(v_1 + v_2) - x_{3j}$, which are $N(0, v_3 + v_1v_2/(v_1 + v_2))$. So the log-likelihood is

$$\begin{aligned}\ln L = &-\frac{p}{2}\ln(2\pi) - \frac{p}{2}\ln(v_1 + v_2)\\ &- \frac{1}{2}\sum_{j=1}^{p}(x_{1j} - x_{2j})^2/(v_1 + v_2)\\ &- \frac{p}{2}\ln(2\pi) - \frac{p}{2}\ln[v_3 + v_1v_2/(v_1 + v_2)]\\ &- \frac{1}{2}\sum_{j=1}^{p}[x_{3j} - (v_2x_{1j} + v_1x_{2j})/\\ &\qquad (v_1 + v_2)]^2/\\ &\qquad [v_3 + v_1v_2/(v_1 + v_2)]. \qquad \text{(A1-1)}\end{aligned}$$

The remainder of the derivation is straightforward after it is realized that

$$\ln(v_1 + v_2) + \ln[v_3 + v_1v_2/(v_1 + v_2)] = \ln(v_1v_2 + v_1v_3 + v_2v_3) \quad \text{(A1-2)}$$

and (dropping the second subscript on the x's for convenience)

$$\frac{(x_1 - x_2)^2}{(v_1 + v_2)} + \frac{[x_3 - (v_2x_1 + v_1x_2)/(v_1 + v_2)]^2}{v_3 + v_1v_2/(v_1 + v_2)} = \{v_3(x_1 - x_2)^2 + [v_1v_2/(v_1 + v_2)](x_1 - x_2)^2 + [v_2^2/(v_1 + v_2)]x_1^2 + [v_1^2/(v_1 + v_2)]x_2^2 - 2[v_1v_2/(v_1 + v_2)]x_1x_2 - 2v_2x_1x_3 - 2v_1x_2x_3\}/[v_1v_2 + v_1v_3 + v_2v_3]. \quad \text{(A1-3)}$$

Collecting terms involving v_1 and v_2, the numerator of the right-hand side of (A1-3) becomes:

$$v_3(x_1 - x_2)^2 + v_2(x_1 - x_3)^2 + v_1(x_2 - x_3)^2$$

after which it is easy to show that (A1-1) is the logarithm of equation (15a).

Appendix 2

REML Solution of the Three-Population Case

Starting from (15a) and taking logarithms, the first derivatives of the log-likelihood are given by

$$\partial \ln L/\partial v_1 = -p(v_2 + v_3)/(2T) - D_{23}^2/(2T) + S(v_2 + v_3)/(2T^2) \quad \text{(A2-1)}$$

$$\partial \ln L/\partial v_2 = -p(v_1 + v_3)/(2T) - D_{13}^2/(2T) + S(v_1 + v_3)/(2T^2) \quad \text{(A2-2)}$$

$$\partial \ln L/\partial v_3 = -p(v_1 + v_2)/(2T) - D_{12}^2/(2T) + S(v_1 + v_2)/(2T^2), \quad \text{(A2-3)}$$

where

$$S = v_1D_{23}^2 + v_2D_{13}^2 + v_3D_{12}^2 \quad \text{(A2-4)}$$

and

$$T = v_1v_2 + v_1v_3 + v_2v_3. \quad \text{(A2-5)}$$

The maximum will lie at the point where all three derivatives are zero. Multiplying the first equation by $2T/(v_2 + v_3)$ we will have

$$-p - D_{23}^2/(v_2 + v_3) + S/T = 0 \quad \text{(A2-6)}$$

and there are similar equations from (A2-2) and (A2-3). Multiplying A2-1 by v_1, A2-2 by v_2, A2-3 by v_3 and adding, we get

$$-2pT/(2T) - S/(2T) + 2ST/(2T^2) = 0 \quad \text{(A2-7)}$$

which yields $S/T = 2p$. From (A2-6) and the two similar equations we then obtain

$$D_{23}^2/p = \hat{v}_2 + \hat{v}_3 \quad \text{(A2-8a)}$$

$$D_{13}^2/p = \hat{v}_1 + \hat{v}_3 \quad \text{(A2-8b)}$$

$$D_{12}^2/p = \hat{v}_1 + \hat{v}_2 \quad \text{(A2-8c)}$$

and these are easily solved to give (16), which can also be written as (17), using equation (15b).