MAXIMUM LIKELIHOOD AND MINIMUM-STEPS METHODS FOR ESTIMATING EVOLUTIONARY TREES FROM DATA ON DISCRETE CHARACTERS

JOSEPH FELSENSTEIN

Abstract

Felsenstein, J. (Department of Genetics SK-50, University of Washington, Seattle, Washington 98195). 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.— The general maximum likelihood approach to the statistical estimation of phylogenies is outlined, for data in which there are a number of discrete states for each character. The details of the maximum likelihood method will depend on the details of the probabilistic model of evolution assumed. There are a very large number of possible models of evolution. For a few of the simpler models, the calculation of the likelihood of an evolutionary tree is outlined. For these models, the maximum likelihood tree will be the same as the "most parsimonious" (or minimum-steps) tree if the probability of change during the evolution of the group is assumed a priori to be very small. However, most sets of data require too many assumed state changes per character to be compatible with this assumption. Farris (1973) has argued that maximum likelihood and parsimony methods are identical under a much less restrictive set of assumptions. It is argued that the present methods are preferable to his, and a counterexample to his argument is presented. An algorithm which enables rapid calculation of the likelihood of a phylogeny is described. [Evolutionary trees: maximum likelihood.]

The first systematic attempt to apply standard statistical inference procedures to the estimation of evolutionary trees was the work of Edwards and Cavalli-Sforza (1964; see also Cavalli-Sforza and Edwards, 1967). At about the same time, the "parsimony" or minimum evolutionary steps method of Camin and Sokal (1965) gave a great impetus to the development of well-defined procedures for obtaining evolutionary trees. Edwards and Cavalli-Sforza concerned themselves with data from continuous variables such as gene frequencies and quantitative characters. The Camin-Sokal approach, on the other hand, was developed for characters which are recorded as a series of discrete states. Although some taxonomists have declared that the problem of guessing phylogenies should be viewed as a problem of statistical inference (Farris, 1967, 1968; Throckmorton, 1968), until recently there have been no attempts to explore the relationship between the statistical inference and minimum-steps approaches. Recently, Farris (1973) has presented a detailed argument that, under certain reasonable assumptions, the maximum-likelihood method of statistical inference appropriate to discrete-character data is precisely the parsimony method of Camin and Sokal. In this paper, I will examine the application of maximum likelihood methods to discrete characters, and will show that parsimony methods are not maximum likelihood methods under the assumptions made by Farris. They are maximum likelihood methods under considerably more restrictive assumptions about evolution.

METHODS OF MAXIMUM LIKELIHOOD

Suppose that we want to estimate the evolutionary tree, T, which is to be specified by the topological form of the tree and the times of branching. We are given a set of data, D, and a model of evolution, M, which incorporates not only the evolutionary processes, but also the processes of sampling by which we obtained the data. This model will usually be probabilistic, involving random events such as changes of the environment, occurrence of favorable

TABLE 1. SOME VARIABLE FEATURES OF A MODEL OF EVOLUTION WITH DISCRETE CHARACTERS.

CHARACTER STATE TREES
Limited to the few states actually observed / Contain further states beyond those observed / Contain states preceding those observed / Contain states preceding and following those observed.
Changes irreversible / Changes reversible.
Known / To be estimated from the data.
PROBABILITIES OF CHANGES
Same for all characters / Different.
Constant per unit time / Constant per segment of the tree / Different in each segment of the tree.
Known / To be estimated from the data.
CHOICE OF CHARACTERS
Random / Only those which have changed at least once in the group / Only those which have changed at least once in some larger reference group / Only those in which all species do not end up having the same state.

mutants, genetic drift, and the sampling of a few individuals from the population by the systematist. We can calculate the probability $P_M(D|T)$ of obtaining the particular set of data (D) given the tree (T) and the model (M). The maximum likelihood estimate of the evolutionary tree is that tree T which yields the largest value of $P_M(D|T)$ for the fixed model M and data D.

Now suppose that we have observed a number of characters across a group. Each character can assume a series of discrete states. In our models, we oversimplify the complex processes of mutation, natural selection, and random genetic drift into sudden changes from one state to another along this character state tree. Even after this oversimplification of the model of evolution, there still seems to be an infinite variety of possible models. Table 1 presents some of the variable features of a model of evolution. There are enough alternatives to result in 768 possible combinations. No pretense is made that Table 1 is exhaustive.

Since it is obviously impossible to discuss all possible models, one of the simpler ones will be discussed as an example. Suppose that the probability of change is known, constant per unit time, and the same for all characters. The choice of characters is random, and all the character state trees are $0\rightarrow1\rightarrow2\rightarrow3\rightarrow4\rightarrow\ \ldots$, so that changes are irreversible. We will calculate the likelihood of the tree A in Figure 1, using this model and the data shown in that figure. The present is called time 0, so that times in the past are negative: this is purely a matter of convention. Let the probability of change be u per unit time. The probability of change during a small interval of time of length dt will be $u\ dt$. We have assumed that the probability of change during any time interval is independent of the times and numbers of previous changes. Thus the probability of k changes during a time interval of length t is the Poisson probability

$$e^{-ut}(ut)^k/k!.$$

Now consider character 1. Since change is irreversible, and species 3 and 4 have state 0, the populations at points 5 and 6 on the tree must also have had state 0. Therefore the segments below points 1, 2, 3, 4, and 5, segments whose lengths in time are 0.96, 0.96, 1.10, 1.10, and 0.14 respectively, must have had 2, 0, 2, 0 and 0 changes. The likelihood of the tree is the probability of the data given the model and the tree, a product of probabilities for individual segments:

$$(1)\quad \begin{aligned} L_1 &= e^{-0.96u}(0.96u)^2(1/2!)e^{-0.96u}e^{-1.10u}\cdot \\ &\quad (1.10u)^2(1/2!)e^{-1.10u}e^{-0.14u} \\ &= e^{-4.26u}(0.96u)^2(1.10u)^2/4. \end{aligned}$$

A directly analogous argument gives the probability for character 2:

$$(2)\quad \begin{aligned} L_2 &= e^{-0.96u}e^{-0.96u}(0.96u)^2(1/2!)\cdot \\ &\quad e^{-1.10u}e^{-1.10u}(1.10u)^2(1/2!)e^{-0.14u} \\ &= e^{-4.26u}(0.96u)^2(1.10u)^2/4, \end{aligned}$$

which happens, in this case, to be the same. Character 3 raises a new problem. It is not immediately apparent what was the state at point 5 on the tree. It could have been any of the states 0, 1, 2, or 3. We can

calculate probabilities associated with each of these alternatives. Since we are interested only in the overall probability of the data given model and tree, we must sum these four probabilities, getting

$$(3)\quad L_3 = e^{-4.26u}(0.96u)^3(0.96u)^3/36 + e^{-4.26u}(0.96u)^2(0.96u)^2(0.14u)/4 + e^{-4.26u}(0.96u)(0.96u)(0.14u)^2/2 + e^{-4.26u}(0.14u)^3/6.$$

Since evolution in the different characters is assumed to be independent, the overall likelihood is the product $L = L_1 L_2 L_3$. We have not so far specified the value of u. If we take $u = 1$, then we can use (1), (2), and (3) to calculate that $L = 1.3344 \times 10^{-8}$ is the likelihood of tree A.

Knowing how to evaluate the likelihood of a tree does not solve all of our problems. We still have to find the maximum likelihood tree. Existing methods for doing this are slightly more sophisticated versions of trial and error. Some initial guess at the topology of the tree is made, and times for the branch points are assigned. We make small changes in these times, evaluating the likelihood after each change. Any time change which results in a higher likelihood is accepted, and the new time becomes the basis for further changes. If a change of time does not result in a higher likelihood, it is rejected, the branch point time being left at its previous value. If changes of times cause two branch points to "collide," it is natural to suspect that the topology ought to be rearranged in that part of the tree. Eventually, following such a procedure, we arrive at a tree which cannot be improved by small alterations. This tree is our guess of the maximum likelihood tree. There is no guarantee that it is the true maximum: all we know is that it is a *local* maximum. Our algorithm is of the "hill-climbing" type, and suffers from the common weakness of all such algorithms—they climb the hill on which the starting point happens to be located, and there is no guarantee that this is the highest hill. One would like some assurance that a hill-climbing algorithm will always give the maximum likelihood tree. If we knew that whenever we started with the wrong tree topology, we would always find that two branch points would collide, then we could have confidence that any topology in which they did not collide, any topology with an "internal" maximum of the likelihood, was the correct topology (although we would still lack assurance that there is only

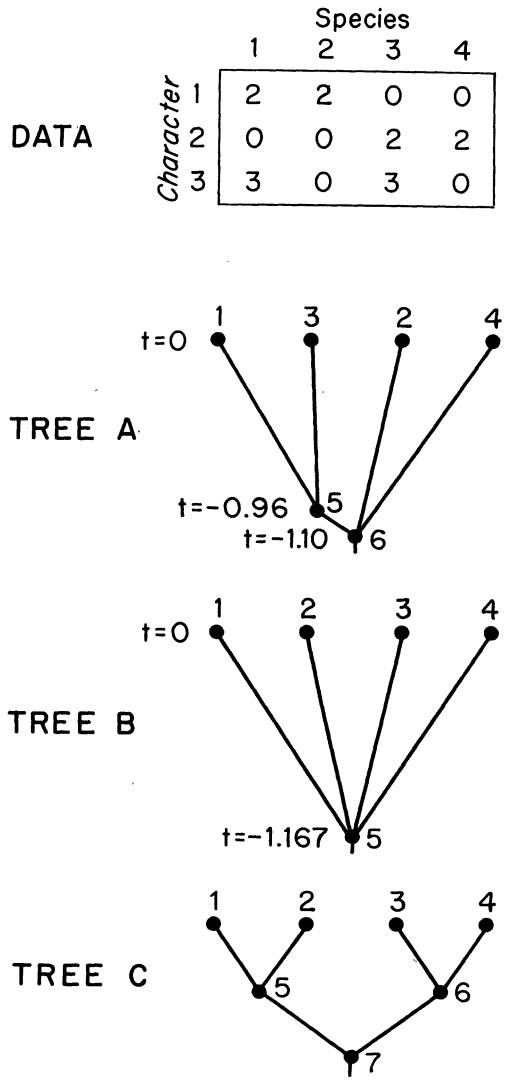

FIG. 1.—A set of data and three possible evolutionary trees, used to illustrate maximum likelihood estimation. See text for details.

one internal maximum for that topology). Another approach would be to develop some means of narrowing down the range of possible trees, making use of the properties of the likelihood function. Estabrook (1968) has done this for the parsimony criterion. Hopefully the same can be done for likelihood.

ESTIMATING ONLY THE TOPOLOGY

The procedures just described result in a maximum likelihood estimate of the tree, which is specified by the topology and the times of the branch points. If we are interested in estimating only the topology, we can use one of two approaches. If we had a probability model for the generation of the tree topology and branch point times, a model incorporating probabilities of speciation and extinction events, we could carry out direct maximum likelihood estimation. This involves calculating the probability of getting the observed data, given the model and the topology of the tree (but not the branch point times). To evaluate the likelihood of a topology τ, we would calculate

$$(4)\quad P_M(D|\tau) = \int_S P[\text{times}|\tau]\, P[\text{Data}|\text{times},\tau],$$

where we integrate over the set S of all branch point times compatible with the topology τ. The analogous procedure for the case of continuous characters is described by Edwards (1970). Even with the largest computers, numerical integration methods would be far too slow to be of any use, so that it would be necessary to evaluate the integrals algebraically. The methods for doing this have not been developed.

An easier approach would be simply to estimate both topology and branch point times, and then to ignore the branch point time estimates. Such a procedure does not make fully efficient use of the data, but it will have to suffice until methods for calculating (4) have been devised. More will be said below about the properties of this second procedure.

ESTIMATING THE PROBABILITY OF CHANGE

So far, we have assumed that u is known in advance. If u is instead to be estimated from the data, a complication arises. Note that in formulas (1), (2), and (3), u always enters into the expression in the form of its product with the length of a tree segment (0.96u, 1.10u, 0.14u, or the sum 4.26u). Likewise, the segment lengths are always multiplied by u. If we halve u and double all the segment lengths, the products will remain the same, and therefore so will the likelihood. In general, u will be completely confounded with the segment lengths, so that only their products can actually be estimated. We might as well assume that $u = 1$, so that the units of "time" we estimate are actually units of expected amount of change per character, ut. A similar situation is encountered in maximum likelihood estimation of evolutionary trees using continuous character data (Felsenstein, 1973). We can place our trees on a true time scale only if u is known and/or if some of our data come from dated fossils.

A SUFFICIENT CONDITION

If it is known in advance that u is very small compared to the evolutionary time elapsed, so that all the ut are very small, we can obtain the maximum likelihood tree without estimating u and without knowing its exact value. It turns out that the maximum likelihood tree is exactly the minimum-steps, or "most parsimonious" tree. To show this we first need an expression for the likelihood of a tree. Let v_i be the length in time of the i-th segment of the tree. For each character there may be many states which could have existed at each branch point in the tree. For example, in equation (3) we had four terms, corresponding to the presence of states 0, 1, 2, and 3 at branch point 5 of tree A. To find the likelihood of a tree, we must sum over all possible combinations of states in the branch points of the tree. Let n_{ijk} be the number of evolutionary steps (state changes) in the i-th character, in the j-th

segment of the tree, according to the k-th possible assignment of character states to branch points. Then the likelihood of the tree is

$$(5)\quad L = \prod_i \sum_k \prod_j e^{-uv_j}(uv_j)^{n_{ijk}}/n_{ijk}!$$

which is a product of Poisson probabilities. This can be rearranged to become

$$(6)\quad L = \prod_i \sum_k u^{\sum_j n_{ijk}} e^{-u\sum_j v_j} \prod_j v_j^{n_{ijk}}/n_{ijk}!.$$

Expressions (5) and (6) are a more general form of the product of (1), (2), and (3).

As $u \to 0$, the individual terms of (6) approach zero, the approach being faster the larger is $\sum_j n_{ijk}$. That quantity is simply the number of steps required in character i over the whole tree, assuming state assignment k. As u becomes small, the term for each character i is contributed almost entirely by that particular pattern of state assignment, k, for which $\sum_j n_{ijk}$ is smallest. Asymptotically, we can drop the summations over k entirely, fixing k at the single value for each character which requires the fewest evolutionary state changes to be assumed. Let the number of state changes required in character i in tree segment j be n_{ij}. Then the likelihood expression becomes asymptotically

$$(7)\quad L \sim u^{\sum_{ij} n_{ij}} \prod_i e^{-u\sum_j v_j} \prod_j v_j^{n_{ij}}/n_{ij}!.$$

Finally, let us take the ratio of the likelihoods of two trees whose minimum total numbers of evolutionary changes are n_1 and n_2. We get

$$(8)\quad \frac{L_1}{L_2} = u^{n_1-n_2}\frac{\prod_i e^{-u\sum_j v_j} \prod_j v_j^{n_{ij}}/n_{ij}!}{\prod_i e^{-u\sum_j v_j'} \prod_j (v_j')^{n_{ij}'}/(n_{ij}')!}.$$

As $u \to 0$ this ratio will approach either zero or infinity, depending on whether $n_1 > n_2$ or $n_1 < n_2$ (if $n_1 = n_2$, it will approach a nonzero constant). Therefore the tree with the highest likelihood has the smallest value of $\sum_{ij} n_{ij}$, so that the most parsimonious tree is also the maximum likelihood tree. This accords with intuition. If we assume that it is a priori very improbable that any evolutionary changes at all will occur, then that tree will strain our credibility least which would require the fewest of these improbable events to explain the observed data.

We now have a sufficient condition for the most parsimonious tree to be the correct maximum likelihood estimate. This would seem to provide us with an acceptable statistical justification for using the parsimony criterion to guess the tree topology. We could then assign the values of the v_i by maximizing the right-hand portion of (7). But there is a fly in the ointment. If our assumption were true that evolutionary change is improbable during the relevant period of time, most characters should be uniform over the group. A few would show a single change of state during the evolution of the group. But only very rarely would we find more than one change of state, so that few or no characters would show convergence. If characters showing no changes were excluded from the data, then each character must have at least one change of state, but very few of the characters would be expected to require more than one change of state. Real data is certainly not like this. It is not unusual to see at least one occurrence of convergence in every character. With each such set of data we encounter, our confidence in the assumption of the improbability of changes should grow less. It therefore seems unlikely that we can justify most uses of parsimony techniques on these grounds.

It might be argued that each particular evolutionary state change is improbable, and that the apparent convergences are merely the result of a taxonomist scoring two distinct states as identical. This assumption does not resolve the dilemma. The problem is not that the second state change occurring in a character is the same as the first. It is the number of state changes, not their identity, which is the problem. Assuming misclassification may

help explain why the second change appears to be a duplicate of the first, but we still see too many cases where more than one state change is required to explain the observed data.

A COUNTEREXAMPLE

If we relax the assumption that the probability of change is small, there is no necessary connection between likelihood and parsimony criteria. This is illustrated by the data and trees in Figure 1. The minimum-steps, or most parsimonious, tree is shown as tree C. It requires a minimum of 10 state changes in evolution. To find trees A and B, I have used an iterative maximum likelihood procedure. The likelihood of each tree examined was calculated using exactly the same arguments which led to equations (1), (2), and (3). A more systematic statement of the procedures used to calculate the likelihoods is given below, in the next-to-last section of this paper. The value of u was assumed to be 1. A simple search algorithm of the "hill-climbing" type was used, altering the times of the branch points one at a time and stopping when no further alteration of any branch point by a given small amount resulted in an improvement of the likelihood of the tree. Since the resulting maximum likelihood estimate seems not to be particularly dependent on the details of the search algorithm, that algorithm will not be described further here. Starting with a tree of form C, and using this sort of maximum likelihood procedure, we arrive at tree B as the tree of highest likelihood. Two of the segments originally present in tree C have shrunk to zero length. The likelihood of tree B is 1.249×10^{-8}. But when we start with a different tree topology, we can obtain tree A, which has a likelihood of 1.3344×10^{-8}, and seems to be the maximum likelihood estimate. It would require a minimum of 11 state changes. This counterexample establishes that there is no necessary identity between parsimony methods and maximum likelihood methods, unless we make assumptions strict enough to exclude the model which leads to likelihoods (1), (2), and (3). The assumptions made by Farris (1973) are not this strict. They easily include the model which leads to those expressions. Yet he is able to argue the identity of parisomy and likelihood estimates, given his assumptions. There is an apparent contradiction here which requires explanation.

THE FARRIS PROCEDURE

Farris makes a very nonrestrictive set of assumptions that easily include the case in Figure 1. He assumes discrete characters, and independence of the number of changes in different characters. He then divides the tree into numerous short segments. He assumes that for each pair of characters i and j, and for each pair of tree segments k and l, the probabilities of a single change of the characters in the segments are nearly enough equal that we always have

$$p_{ik} > p_{jl}^2, \text{ for all } k \text{ and } l,$$

whenever i = j or whenever the two characters i and j are ones which turn out to be incompatible in the sense of Camin and Sokal (1965). Farris also assumes that if $p_{ik}^{(n)}$ is the probability of n character-state changes in character i in segment k, then

$$p_{ik}^{(n)} < [p_{il}^{(1)}]^n$$

for all i, k, and l, and for $n > 1$.

If the characters have equal probabilities of change per unit time, and if the tree is divided into segments of equal length, the first set of requirements will always be met. If the probability of k changes in a segment is also the Poisson probability

$$e^{-ut}(ut)^k/k!,$$

then the second set of requirements will be met provided the segments are short enough that $ut < \log_e 2$. In the counterexample given in the preceding section, the characters had equal probabilities of change per unit time. We therefore expect Farris' arguments to apply to those cases

if we divide the tree into short segments of equal length.

Farris starts by estimating not only the topology and the times of branching, but also the phenotypes in the populations at the end of each of these short time intervals. Thus he is estimating not only the tree but the "pathways followed by evolving populations through time and through character space." For those readers who have trouble visualizing this, Figure 2 pictures one of these entities, compatible with tree C in Figure 1. Farris shows convincingly that the pathway which has the highest likelihood has the smallest number of steps required. To make a maximum likelihood estimate of the topology of the tree, he discards all the information he has estimated, except for the tree topology. He discards not only the host of estimated phenotypes, but the branch point times as well. The number of state changes needed on a tree depends only on the tree topology. Finding the "pathway" with the highest likelihood, then discarding all information but the topology, one will arrive at the minimum-steps topology.

The problem with his argument is that the procedure I presented earlier can be used in the same way, and gives different results. Taking tree A from Figure 1 and dropping the times of branching, we arrive at an estimate of the topology which is not the same as tree C, the minimum-steps topology. As I mentioned above, neither of these two procedures gives a true maximum likelihood estimate of the tree topology. To get such an estimate, we would have to maximize P[data|topology]. The maximum likelihood tree topology could be searched for, using (4)—if we could evaluate that likelihood expression. Both Farris' and my estimates of the tree topology are obtained by first estimating more than the topology, then dropping some of that information. This is not the same as making a maximum likelihood estimate of the topology. If it were, our two estimates of the topology should always

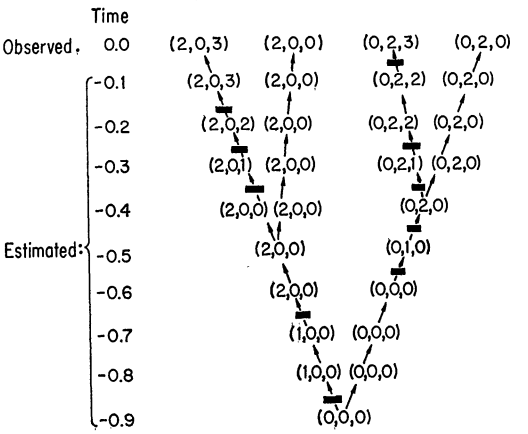

FIG. 2.—A "pathway" of the sort estimated by Farris (1973). Data are those given in Figure 1. Each triple represents the phenotypes of the three characters in one population. Steps are marked by bars across the arrows.

be identical, which they are not. Only the expression based on (4) is the maximum likelihood estimate of the topology. If we cannot use (4), either because we have no model of branching to give us P[times|τ] or because we cannot evaluate the integrals, my procedure would seem to have at least one major advantage, consistency (in the statistical sense).

CONSISTENCY OF ESTIMATES

An estimate has the property of *consistency* if, as we sample more and more data, the estimate converges on the true value. Maximum likelihood estimates have this property under a wide variety of circumstances (Wald, 1949). Wald gives eight conditions which, if all are satisfied, guarantee that the maximum likelihood estimate is consistent. These conditions are too complex to discuss here in detail, but it can be shown that estimates based on the likelihood expressions (4), (5), and (6) satisfy them, so that these estimates are consistent. Expression (4) is used to estimate the tree topology, expressions (5) and (6) are used to estimate the tree, including the branch point times. If we estimate the tree topology by first estimating the tree, then dis-

carding the branch point times, we are not carrying out maximum likelihood estimation of the tree topology. However, the estimate of tree topology will nevertheless be consistent. As we collect more and more data, the estimate of the tree converges on the true tree, with respect to both its topology and its branch point times. A forteriori, then, the estimate of the topology obtained by ignoring the branch point times must converge on the true topology.

This will not necessarily apply to Farris' estimate of the topology. First one would have to prove the consistency of the estimate of the "pathway." But in estimating the "pathway," the number of parameters being estimated increases without limit as the number of characters increases. Each new character brings with it a host of states to be estimated. The assumptions of Wald (1949) are violated in such a case. Wald assumes that there are a finite number, k, of unknown parameters to be estimated. He also assumes that the observed variables are independent and identically distributed. In the present case the observed variable is the n-tuple of observed tip phenotypes for one character. If the distribution of each character depends on different unknown paramenters (the node phenotypes and the initial character state), Wald's assumption of identical distributions is violated. In fact, any likelihood method which estimates even one ancestor's state per character will not necessarily give a consistent estimate. Of course, nothing said here rules out the possibility that Farris' estimate of tree topology has the property of consistency. Wald's conditions are sufficient, but they are not necessary for consistency. For some examples and discussion of the problem of infinitely many "nuisance parameters," the reader is referred to the paper by Kalbfleisch and Sprott (1970).

AN ALGORITHM FOR CALCULATING LIKELIHOODS

Expressions (1), (2), and (3) were derived by inspection, and equations (5) and (6) involve summing over all possible assignments k of states to branch points, a very large number of possibilities. It may be of interest to derive a systematic procedure for calculating the likelihood which is applicable in a great many cases and which uses a common property of many models to greatly simplify the calculation. In many models, the probability of changing from state i to state j during one segment of the tree does not depend in any way on how the population arrived at state i in the first place. Therefore the changes of state constitute a Markov process. Since we assume independence of evolution in different characters, we calculate the likelihood of the tree separately for each character and then multiply these. Therefore we need only consider how to calculate the likelihood for one character.

Consider two nodes (tips or branch points), i and j, which have the same immediate ancestor, k. Let L_{im} be the probability of obtaining the data observed on all the tips above node i, given that node i has state m. Suppose that we are given the L_{im} for all values of m, and for node j we are given the L_{jn} for all values of n. We want to calculate for their ancestor, node k, the L_{kp} for all values of p. We are given $P_i(p,m)$ and $P_j(p,n)$, the transition probabilities for segments k – i and k – j of the tree. That is, $P_i(p,m)$ is the probability that in segment k – i we end up with state m given that we started with state p.

To calculate the L_{kp} we use

$$(9)\quad L_{kp} = \sum_m \sum_n P\,[p \to m \text{ in segment } k-i \textit{ and } p \to n \text{ in segment } k-j]\, L_{im} L_{jn}$$
$$= \sum_m \sum_n P_i(p,m)\, P_j(p,n)\, L_{im} L_{jn}$$
$$= \left(\sum_m P_i(p,m)\, L_{im}\right)\left(\sum_n P_j(p,n)\, L_{jn}\right).$$

The indices m, n, and p run over all states of the character.

Of course, to use this formula we must know the $P_i(p,m)$. These will depend on the specific model of evolution assumed for the character. In the example given earlier in this paper, the character state

tree was of the form $0 \rightarrow 1 \rightarrow 2 \rightarrow \ldots$, so that

$$(10)\quad P_i(p,m) = e^{-u(t_i - t_k)} \times [u(t_i - t_k)]^{m-p}/(m-p)!$$

where $t_i - t_k$ is the duration of segment k – i. If the character state tree, and the probability of change, u, are the same for all characters, we can use the same transition probabilities $P_i(p,m)$ for all characters. The transition probability from state p to state m will differ for each segment, since the segment lengths will differ.

We start the calculation of the likelihood by assigning values of L_{im} to the tips of the tree. Since the states of the tip populations have (usually) been observed, one of the L_{im} will be 1, and the rest will be zero. If the state of the character in a particular tip is not known, this can be taken into account by setting *all* the L_{im} to 1 for that tip (recall that L_{im} is the *probability of obtaining the observed data* at or above node i, given that the state at that point is m). Another possible complication is when the observed phenotype is compatible with a number of underlying states. In that case, the L_{im} corresponding to those states are 1 and the rest are zero. This will be the case with protein sequence data, where the observations are amino acids, each of which is compatible with a set of underlying codons. We will have 64 states, and 64 × 64 transition probabilities for each tree segment. The observed amino acids each define the set of possible codons at a tip.

Given the L_{im} for each tip, we now proceed down the tree. There will always be at least one branch point which is the immediate ancestor of two or more tips. We can use (9) to calculate the L_{kp} for that branch point. Now consider the two tips to have been "pruned" off the tree, and consider the branch point to have become a tip. We once again have a tree with values of L_{im} for each tip (but now there is one less tip on the tree). We can repeat the process, calculating the L_{im} for another branch point, pruning off the segments above that node, and so on. We continue the process until we arrive at the bottom branch point and have calculated its L_{im}. If the state of this original population is known to have been, say, state s, the likelihood of the tree with respect to the character is simply L_{is}, by definition. If we did not know state s in advance, one might think that we could estimate it by choosing the largest of these final L_{im}. But this introduces one new parameter to be estimated per character. As the number of characters grows by collection of new data, so will the number of parameters being estimated, and we may lose the property of consistency of the tree estimate.

In certain cases, we should be able to get around this. If the character state tree contains an infinite number of states both preceding and following the ones observed, the character state tree is $\ldots \rightarrow -2 \rightarrow -1 \rightarrow 0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow \ldots$. Then when we observe states 0 and 3, these numbers have no absolute meaning: all we know is that they are 3 states apart on the tree. If we had observed states 1 and 4, we might have coded them as states 0 and 3. If the true initial state is 0, the probability of observing two states three steps apart is

$$P(0,3|0) + P(1,4|0) + P(2,5|0) + \ldots,$$

where $P(i,j|k)$ is the probability of observing states i and j given that the initial state in the original population was k. By the symmetry of the situation, this is equal to

$$P(0,3|0) + P(0,3|-1) + P(0,3|-2) + \ldots.$$

Thus if all we can observe are the differences between states and we get no information from the absolute identities of the states, we can obtain the likelihood of these differences by calculating for the bottom branch point of the tree

$$(11)\quad L = \sum_m L_{im}.$$

This gets us around the necessity of estimating one new parameter for each character. It is a "marginal likelihood" procedure

as defined by Kalbfleisch and Sprott (1970). In the case of continuous characters, a closely analogous procedure has been derived (Felsenstein, 1973). Of course, this procedure will not work if we do not have this kind of character state tree: how to deal with this problem more generally is not clear.

PERSPECTIVE

Maximum likelihood and "parsimony" are not, in general, identical. The conditions for the parsimony method to be justifiable as the maximum-likelihood method are often not met. Although it is possible to carry out maximum likelihood estimation directly, the computing procedures for doing so are rather slow. Furthermore, it should be emphasized that it is rather difficult to find cases in which the parsimony and maximum likelihood methods give different results. For many sets of data, the "parsimony" method may be a good approximation to the maximum likelihood method even when probabilities of change of the character are not small. However, if the parsimony technique is taken to yield a genuine maximum likelihood estimate, the assumptions which are thought to make this interpretation possible should be clearly stated. If evolutionary trees are to be inferred in any justifiable way, then the maximum likelihood criterion (or some other statistical inference method) must be used to infer them.

ACKNOWLEDGMENTS

I have benefited greatly from discussions with A. W. F. Edwards, J. S. Farris, L. H. Throckmorton, G. F. Estabrook, and E. A. Thompson. This paper is based in part on a dissertation in candidacy for the Ph.D. degree in the Department of Zoology, University of Chicago. I wish to thank my thesis adviser, R. C. Lewontin, especially for his patience. That part of the work was supported by NIH Genetics Training Grant No. 5T01 GM-00090. Subsequent work ancestral to this paper has been supported by Postdoctoral Research Fellowship 1-F2-GM-36,536-01 from the National Institute of General Medical Sciences, by U.S. Atomic Energy Commission Contract AT (45-1) 2065, and by Task Agreement Number 5 of U.S. Atomic Energy Commission Contract AT(45-1) 2225, the latter two with the University of Washington.

REFERENCES

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

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 32:550–570 (also in Amer. J. Human Genetics 19:233–257).

EDWARDS, A. W. F., AND L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees. pp. 67–76 *in* Phenetic and phylogenetic classification, eds. V. H. Heywood and J. McNeill. Systematics Association Publication No. 6, London.

EDWARDS, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. J. Royal Statistical Soc. B 32:154–174.

ESTABROOK, G. F. 1968. A general solution in partial orders for the Camin-Sokal model in phylogeny. J. Theoretical Biol. 21:421–438.

FARRIS, J. S. 1967. The meaning of relationship and taxonomic procedure. Syst. Zool. 16: 44–51.

FARRIS, J. S. 1968. Categorical ranks and evolutionary taxa in numerical taxonomy. Syst. Zool. 17:151–159.

FARRIS, J. S. 1973. On the use of the parsimony criterion for inferring evolutionary trees. Syst. Zool. 22:250–256.

FELSENSTEIN, J. 1973. Maximum likelihood estimation of evolutionary trees from continuous characters. Amer. J. Human Genetics (in press).

KALBFLEISCH, J. D., AND D. A. SPROTT. 1970. Application of likelihood methods to models involving large numbers of parameters. J. Royal Statistical Soc. B 32:175–208.

THROCKMORTON, L. H. 1968. Concordance and discordance of taxonomic characters in *Drosophila* classification. Syst. Zool. 17:355–387.

WALD, A. 1949. Note on the consistency of the maximum likelihood estimate. Annals of Mathematical Statistics 20:595–601.

Manuscript received March, 1973

Note added in proof: Since this paper was submitted, Farris has revised his procedure, apparently generalizing it. If I understand his new procedure correctly, the critique given above continues to apply.