A Probability Model for Inferring Evolutionary Trees

James S. Farris

Systematic Zoology, Vol. 22, No. 3. (Sep., 1973), pp. 250-256.

Stable URL:
http://links.jstor.org/sici?sici=0039-7989%28197309%2922%3A3%3C250%3AAPMFIE%3E2.0.CO%3B2-W

Systematic Zoology is currently published by Society of Systematic Biologists.

A PROBABILITY MODEL FOR INFERRING EVOLUTIONARY TREES[1, 2]

JAMES S. FARRIS

Abstract

Farris, J. S. (Dept. of Ecology and Evolution, State Univ. of New York at Stony Brook, Stony Brook, N.Y. 11790) 1973. A Probability Model for Inferring Evolutionary Trees. Syst. Zool. 22:250–256.—Estimation of evolutionary trees should be treated as a problem in statistical inference, but such treatment requires the explicit formulation of a stochastic model of the evolutionary process. Because an evolutionary inference procedure is likely to be put to such uses as deciding the issue of whether rates of evolution are homogeneous, the stochastic model underlying the inference procedure should not assume homogeneity over time of the evolutionary process, and in fact, should make only the weakest evolutionary assumptions necessary. Such a model is constructed, and it is shown that most parsimonious trees are maximum-likelihood estimated evolutionary trees under the stochastic model. Similarity clustered phenograms appear not to be well justified as statistical estimates of evolutionary trees, even when homogeneity of evolutionary rates is assumed. [Evolution; trees; maximum-likelihood; phenograms.]

While it is generally agreed that the reconstruction of evolutionary trees should ideally be regarded as a problem in statistical inference, few approaches to evolutionary taxonomy have taken into account the full implications of that premise. A statistical inference procedure can properly exist only as a method derived under a specified probability model and demonstrably possessing one or more optimality properties under that model. Stochastic models of the evolutionary process have seldom been discussed in the context of evolutionary inference problems, and explicit consideration of the statistical optimality properties of evolutionary inference methods has consequently been neglected. The purpose of the present paper is to construct a simple probability model of the evolutionary process and to discuss the desirability of some inference methods under that model.

SELECTION OF AN OPTIMALITY CRITERION

Because the problem under consideration is that of inferring evolutionary trees, many familiar statistical optimality criteria must be left unused. While such properties as unbiasedness and minimum-variance are well justified and easily applied in the case of normally distributed, real valued, random variables, it is not clear how they could be applied to—or even defined for—estimated trees. The only statistical optimality principle which would seem readily applicable to the case of estimated trees is that the selected tree should be the most probable tree on the basis of available data. For some models, we may be able directly to formulate a function $P\{E|D\}$, the conditional probability of an evolutionary hypothesis, E, given specified data, D. For other models it may be more natural to specify the function $P\{D|E\}$, the probability of data given a specified evolutionary hypothesis, then to obtain $P\{E|D\}$ through the inversion formula

$$P\{E|D\} = \frac{P\{D|E\}P\{E\}}{P\{D\}}. \qquad (1)$$

In either case E is selected to maximize $P\{E|D\}$ for a specified D. For this purpose, it is necessary only to know the rank ordering of $P\{E|D\}$ as a function of E for fixed D. This simplification is frequently important, for example where $P\{E|D\}$ is obtained through the inversion formula (1),

[1] Preparation of this paper was partially supported by NSF Grant B036060.

[2] Contribution No. 62 from the Department of Ecology and Evolution, State University of New York at Stony Brook.

P{D}, being constant for fixed D, can be ignored. In the same case P{E}, the probability of evolutionary hypothesis E not conditional upon any data, may be treated as if equal for all E. The evolutionary hypothesis, E, that maximizes P{E|D} for a specified D is called the *maximum-likelihood* estimate of the true evolutionary tree. A general discussion of maximum-likelihood estimation is provided by Lindgren (1962), which will also serve as a reference for other probability-theoretic results in this paper.

SELECTION OF A STOCHASTIC MODEL

Given that we select estimated evolutionary trees always according to the maximum-likelihood criterion, the method for constructing an estimated evolutionary tree is in principle well determined once a stochastic model of the evolutionary process has been selected. This model should be chosen with some care because of the spectrum of uses to which a method for inferring evolutionary trees is likely to be put. For example, there has been some controversy on the issue of whether rates of evolution are homogeneous over time and among phyletic lines, particularly in the case of protein sequences. An attractive way of collecting evidence relevant to this controversy is to construct an evolutionary tree and to inspect it for indications of the truth or falsity of the hypothesis of rate homogeneity. But if a tree-inference procedure is to be used in this way, it is plainly undesirable for the procedure to impose an artificial homogeneity on the estimated evolutionary rates. Hence the stochastic evolutionary model should be selected so that it does not assume the evolutionary process to be homogeneous over time.

The evolutionary inference model of Cavalli-Sforza and Edwards (1967) will provide an instructive example of an evolutionary model that does assume homogeneity over time. Here the data are taken to consist of allelic frequencies at each of several loci for several populations, and evolutionary change in these frequencies is taken to be due entirely to genetic drift. The estimated tree is characterized by its shape and the placement on its branches of the data populations, together with the estimated amount of evolution occurring in each branch of the tree and the time (in relative units) of each branching point of the tree. Let t_i denote the temporal length of the ith branch of a tree and d_i the amount of evolution in the ith branch. Cavalli-Sforza and Edwards derive a frequency function $f(d_i|t_i)$ giving the density of the event that d_i units of divergence will occur in a branch of temporal length t_i under the genetic drift model. The density of an entire tree is then

$$f_E = \prod_i f(d_i|t_i). \qquad (2)$$

Provided the tree connects all the data populations in both time and gene-frequency space, equation (2) gives the relative frequencies of trees E for specified data, and so is the continuous analog of P{E|D}.

Formulations such as that of Cavalli-Sforza and Edwards are readily enough constructed for a variety of stochastic evolutionary models, and they have the appeal of providing estimates of the times of divergence of phyletic lines leading to modern species. This type of approach, however, suffers from two major drawbacks. Frequency functions of the type $f(d_i|t_i)$ can be formulated for simple models only under the premise that the evolutionary process is homogeneous over time, inducing, as noted above, an undesirable restriction on the applicability of a method. Also, even in the case where a relatively simple, time-wise homogeneous process is postulated, frequency functions of the form of (2) may be quite intractable. The maximum-likelihood estimated tree under the procedure of Cavalli-Sforza and Edwards, for example, must be identified by minimizing

$$\sum_i (d_i/t_i + \log t_i) \qquad (3)$$

where summation extends over all the branches in the estimated tree. While nat-

urally it would be possible to construct a function $f(d_i|t_i)$ incorporating some specific sort of heterogeneity of the stochastic process over time, there seems to be little biological motivation for choosing any particular such function, and every reason to believe that any specific choice of such a function would result in an inference procedure quite unsuited to the majority of applications.

It seems desirable to avoid the assumption of stochastic homogeneity of the evolutionary process, and consequently to avoid formulations of the sort used by Cavalli-Sforza and Edwards, in which a distribution of amount of change as a function of time must be specified. More generally, it would seem that the safest course to follow in constructing a stochastic model of the evolutionary process for purposes of deciding upon inference procedures is to keep the evolutionary assumption made in the model as simple and weak as possible so as to obtain an inference procedure applicable to a variety of types of comparative data.

AN EVOLUTIONARY MODEL

To construct a simple stochastic model of evolution, I shall initially consider discrete-valued characters. The generalization of the model to continuous characters will be treated below. Evolutionary modifications in discrete characters will be taken to occur in units of "changes." I shall assume that each species is characterized by exactly one state for each character in the study, and that the differences in number of changes between any two states of any character is known. Note that this last admits of a variety of types of comparative data. For example, if a character comprises identity of the amino acid at a specified site of a protein, we might define the number of changes between any two states, or amino acids, to be unity, or alternatively, as equal to the minimum mutation distance between the amino acids in the sense of Fitch and Margoliash (1967). If the states of the character are coded integer values, we could take the number of changes between two states to be equal to the absolute value of the difference between their numeric codes. In general, a character may be characterized by a state by state matrix of distances expressed in number of changes. I shall assume evolutionary events occurring at different times or in different parts of the evolutionary tree to be probabilistically independent and different characters to be probabilistically independent with respect to their evolutionary changes. I shall *not* assume that characters are restricted to evolve in only one direction.

Let the total temporal length of the evolutionary tree be fixed, while unknown, and to be equal to n in some time units. Divide the branches of the evolutionary tree each into several small time units of length u so that the total number of time units is

$$N = n/u. \tag{4}$$

Since the characters are discrete it seems natural to take a Poisson-type probability model. Therefore, suppose that for u sufficiently small, the probability that any character changes more than once in any time interval is negligible. Denote as p_{ij} the probability that the ith character changes in the jth of N time intervals. Define

$$s_i = \sum_{j=1}^{N} p_{ij} \tag{5}$$

and

$$M_i = \max_j \{p_{ij}\}. \tag{6}$$

Suppose that for sufficiently small u, s_i is independent of u for every character; that is, s_i tends to a well defined limit as u tends to zero. Finally, assume that M_i tends to zero as u tends to zero and that for any pair p_{ij}, p_{kl}, p_{ij}/p_{kl} tends to a finite limit as u tends to zero.

Biologically, these restrictions on the behavior of p_{ij} seem realistic in that even a character that has high probability of changing repeatedly during the evolution-

ary process is unlikely to change during a particular time unit, provided that the time units are sufficiently small and numerous. As each small time unit has its own probability, p_{ij}, of a change's occurring in each character, the evolutionary process is plainly not restricted to be homogeneous over time. Since for sufficiently small u the probability of more than one change within a single time unit in any character is negligible, the probability, p_{ij}, that the ith character changes in the jth time unit is also the *local* mean rate of evolution for the ith character in the jth time unit. The quantity s_i represents the expected number of changes in the ith character over the entire evolutionary tree; the restriction that s_i tend to a well defined limit as u tends to zero simply implies that this expectation does not depend upon an arbitrary choice of time intervals. The restriction that there be a finite limit for every p_{ij}/p_{kl} as u becomes small merely requires that the local average rates of evolution of the characters may vary over time and among characters by any "reasonable" (that is, non-infinite) amount.

Now define d_{ji} to be a variable that takes on value unity if the ith character changes in the jth time unit, and is zero otherwise. Then we can write the probability of a particular sequence of changes on an evolutionary tree as

$$P\{E|D\} = \prod_i \prod_j (p_{ij}d_{ij} + (1-p_{ij})(1-d_{ij})), \quad (7)$$

where we suppose that the sequence of changes under consideration is so restricted that the data species are contacted by the branches of tree E in the character space defined by the available data variables.

Now provided that

$$\max_{g,h,i,j} \{p_{gh}p_{ij}\} < \min_{w,x,y,z} \{p_{wx}(1-p_{yz})\} \quad (8)$$

then $P\{E|D\}$ is monotone decreasing on

$$L(E|D) = \sum_i \sum_j d_{ij}, \quad (9)$$

where by the notation $L(E|D)$ the restriction of E to conform to D is explicitly recognized. In the case that E does not conform to D, formally define $L(E|D)$ to be unbounded and positive. $L(E|D)$ is monotonically related to $P\{E|D\}$ under inequality (8), since (8) specifies that replacing any two changes by any single change must increase the probability of the sequence of changes. As u tends to zero, condition (8) is automatically fulfilled, for then $1 - p_{yz}$ tends to unity, since M_y tends to zero as u tends to zero. Define

$$M^* = \max_i \{M_i\};$$
$$m^* = \min_{i,j} \{p_{ij}\}. \quad (10)$$

Relation (8) is satisfied provided

$$(M^*/m^*)M^* < 1. \quad (11)$$

This must hold, for M^*/m^* must be bounded as the variation among the p_{ij} is bounded, and M^* itself tends to zero as u tends to zero, since it is the maximum of quantities that individually tend to zero for sufficiently small u.

We will usually be content to infer just the shape of an evolutionary tree, rather than the complete information on sequences of changes subsumed by an evolutionary hypothesis as described above. We can do this by selecting as our estimated tree shape, the tree shape T having the maximum possible value of $P\{E|D\}$, for E any evolutionary hypothesis having tree shape T. Because $P\{E|D\}$ is monotone decreasing on $L(E|D)$ under the present stochastic model, we may identify the optimum estimated T by finding a tree shape T with a minimum value of $L(T|D)$:

$$L(T|D) = \min_{E \in A(T)} \{L(E|D)\}, \quad (12)$$

where $A(T)$ is the collection of all evolutionary hypotheses, E, having tree shape T. The problem of searching for the optimum estimated tree shape T is now greatly simplified, for $L(T|D)$ is the "number of steps" for tree T given OTU's and characters D in the sense of Camin and Sokal (1965). Algorithms for calculating $L(T|D)$ are already known: the method of Farris

(1970a) is applicable to real-valued characters, while the method of Fitch (1971) can be used to determine L(T|D) for amino acid sequences. Hence we may identify the tree shape T with minimum value of L(T|D) directly from the data, there being no need to explicitly formulate any of the complete evolutionary hypotheses E.

The procedure derived above permits the construction of an estimated tree shape T by establishing the shape of the evolutionary tree corresponding to the most probable evolutionary hypothesis E on the basis of given data. It would be mathematically more pleasing to select the estimated tree shape T directly by choosing T to maximize

$$P\{T|D\} = \int_{A(T)} P\{E|D\}. \qquad (13)$$

The evaluation of the integral of expression (13), however, requires, as far as I have been able to determine, the imposition of additional assumptions upon the stochastic evolutionary model. Hence in the interest of retaining as much generality as possible in the inference procedure, it seems preferable to select as the best estimated T the tree shape corresponding to the hypothesis E of maximum probability given the data, rather than to try to judge the probability of T directly on the basis of the data.

The extension of this model to the case of continuous characters is readily enough accomplished. If, as has recently been suggested by Eldredge and Gould (1972), continuous-valued characters typically evolve in short, rapid bursts of change separated by periods of very little change, then the evolution of a "continuous" character may be reasonably approximated by a discrete variable. In this case, we will require some scaling factor to convert differences among states of a continuous character into numbers of "changes." If the average size of a "change" in a continuous character is proportional to the within-population variability of that character, then the desired transformation can be accomplished by normalizing each continuous character according to its within-population variability. If X denotes a continuous character with average within-population standard deviation σ, the normalized equivalent of X would be

$$X^* = X/\sigma. \qquad (14)$$

If each of several continuous variables has about the same average probability of changing during a small time interval, transformation (14) is justified only if the transformed variables show about the same average rates of evolution. The theoretical argument of Farris (1966) and the empirical findings of Farris (1970) and of Kluge and Kerfoot (1973) that rates of evolution in continuous characters are well correlated with the within-population standard deviations of those characters suggests that the latter is true, and hence that transformation (14) is an appropriate means of treating continuous characters under the present stochastic model.

EVALUATION OF TREE-CONSTRUCTING METHODS

The results developed above bear on the desirability of two general types of existing methods for inferring evolutionary trees: most parsimonious trees and similarity clustering.

Since L(T|D) is, as noted above, the "number of steps" of a tree shape T in the sense of Camin and Sokal (1965), it is the criterion to be minimized both under the parsimony criterion as originally suggested by Camin and Sokal and the maximum-likelihood approach suggested here. Under the model constructed above, therefore, most parsimonious trees are also maximum-likelihood-estimated trees. While this result seems intuitively reasonable, there had been some doubt of it in the past. It has occasionally been commented that most parsimonious trees are good estimates of the true evolutionary tree only under the assumption that evolution proceeds parsimoniously: that only one change in a character during the entire evolutionary sequence is more likely to occur than any

larger number of changes, or that cases of parallelism are rare in nature (see, for example, Jardine and Sibson, 1971). In terms of the model presented above this "justification" corresponds to the restriction $s_i < 1$ for every character. No such restriction, however, is necessary to justify the choice of the tree shape T with minimum value of $L(T|D)$ under the development above. Suppose, for example, that the data consist of two-state variables, for each of which $s_i = 25$, say. Then the expected number of extra steps is 24 per character, but the tree with minimum total length is still the maximum-likelihood-estimated tree. Hence criticisms of the parsimony criterion of the sort described here would appear largely injustified.

Phenetic similarity-clustering is usually justified as a means of evolutionary inference through the assumption of homogeneity of evolutionary rates. If all the probabilities p_{ij} are assumed to be equal, then the general probability model developed above becomes a model postulating homogeneity of evolutionary rates over time and among phyletic lines. The remainder of the model is unchanged, so that the tree shape T with minimum value of $L(T|D)$ is still the maximum-likelihood estimate of the true evolutionary tree. Since most parsimonious trees need not have the same shape as trees produced by phenetic similarity clustering for the same data, it would seem that phenetic similarity clustering is poorly justified as a way of inferring evolutionary trees, even under the assumption of stochastic homogeneity of evolutionary rates. While this observation appears frequently to have been overlooked, it is not new: the Cavalli-Sforza and Edwards model described previously provides another example of a maximum-likelihood inference procedure in which the stochastic model assumes homogeneity of evolutionary rates and for which the solution trees need not have the same form as similarity-clustered phenograms based on the same data.

Colless (1970) and Goodman and Moore (1971) have constructed justifications of phenetic similarity clustering as a means of evolutionary inference depending not on homogeneity of evolutionary rates, but on homogeneity of rates of net divergence among species. The latter assumption would seem open to criticism on biological grounds. While it is easy to imagine mechanisms which could result in homogeneity in rates of evolutionary change, as for example genetic drift, it is much more difficult to construct models under which the net rates of divergence among species are homogeneous. If rates of evolution are taken to be nonhomogeneous, then rates of divergence can be made homogeneous only by appropriate selection of a pattern of convergence among species. This latter would seem unrealistic, as it would require "coordination" of the evolutionary activities of separately evolving phyletic lines, which might, after all, reside on different continents. Hence, we may realistically believe in homogeneity of rates of net divergence among OTU's only in the case where we believe rates of evolution also to be homogeneous. But if evolutionary rates are assumed to be homogeneous, we have already seen that the maximum-likelihood estimated evolutionary tree need not be the same as the similarity-clustered phenogram. Thus it would appear that in no case can a similarity-clustered phenogram be statistically justified as a means of inferring evolutionary relationships.

Most parsimonious trees are justified as maximum-likelihood estimates of the true evolutionary tree under the rather weak assumptions of the stochastic model developed above. Similarity-clustered phenograms, on the other hand, do not appear to correspond to maximum-likelihood estimated trees under any stochastic model of the evolutionary process. It would therefore appear that most parsimonious trees are preferable to similarity-clustered phenograms for purposes of evolutionary inference.

ACKNOWLEDGMENTS

Dr. J. A. Hartigan, Mr. George Estabrook, and Dr. J. Felsenstein have contributed many useful comments on previous versions of this manuscript. Mrs. Ethel Savarese has performed invaluable secretarial service in the typing of the manuscript.

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 21:550–570.

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

ELDREDGE, N., AND S. J. GOULD. 1972. Punctuated equilibria: An alternative to phyletic gradualism. *In* Models in Paleobiology. T. J. M. Schopf (ed.). Freeman, Cooper & Co., San Francisco. pp. 82–115.

FARRIS, J. S. 1966. Estimation of conservatism by constancy within biological populations. Evolution 20:587–591.

FARRIS, J. S. 1970. On the relationship between variation and conservatism. Evolution 24:825–827.

FARRIS, J. S. 1970a. Methods for computing Wagner trees. Syst. Zool. 19:83–92.

FITCH, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. 20:406–416.

FITCH, W. M., AND E. MARGOLIASH. 1967. The construction of phylogenetic trees. Science 155:279–284.

GOODMAN, M., AND G. W. MOORE. 1971. Immunodiffusion systematics of the primates. I. The Catarrhini. Syst. Zool. 20:19–62.

JARDINE, N., AND R. SIBSON. 1971. Mathematical Taxonomy. Wiley, London 286 p.

KLUGE, A. G., AND W. C. KERFOOT. 1973. The predictability and regularity of character divergence. Amer. Nat. 107:426–442.

LINDGREN, B. W. 1962. Statistical Theory. Macmillan, New York 427 pp.

Manuscript received September, 1972
Revised May, 1973