THE UNIVERSITY OF CHICAGO

STATISTICAL INFERENCE AND THE ESTIMATION OF PHYLOGENIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ZOOLOGY

BY
JOSEPH FELSENSTEIN

CHICAGO, ILLINOIS
MARCH, 1968

"I am not a detective," Major Danby replied with indignation, his cheeks flushing again. "I'm a university professor with a highly developed sense of right and wrong, and I wouldn't try to deceive you. I wouldn't lie to anyone."

"What would you do if one of the men in the group asked you about this conversation?"

"I would lie to him."

Joseph Heller, Catch-22

ACKNOWLEDGMENTS

I would like to thank Dr. R. C. Lewontin for his advice and encouragement. I would also like to thank Dr. Lynn Throckmorton for many helpful conversations. The basic idea of this thesis was suggested by Dr. A. W. F. Edwards, in a conversation, for which I am indebted to him. I have also greatly benefited from discussions with James Farris, Glenn Sharrock, and Harold Voris. None of the above are responsible for any errors or misstatements.

I am grateful to my father, Jacob Felsenstein, for drawing the figures for this thesis.

Dedicated with respect to Jesse Lemisch, Richard Levins, Donald Barnett, Staughton Lynd, and the fifty-eight who stayed.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

INTRODUCTION

In the last ten years there has been an increasing controversy over taxonomic methods. This has in part been stimulated by the availability of electronic computers and the possibility of analyzing taxonomic data automatically. The primary controversy has been over whether taxonomic classification should be based on phylogenetic relationship or phenotypic (phenetic) similarity. The purpose of a hierarchical classification is to facilitate the storage and retrieval of information concerning the organisms. The controversy is thus as much technological as biological. All parties admit that a real phylogeny underlies observed biological similarities. Methods for producing phylogenies can be discussed in terms of the truth or falsity of the results, in contrast to methods for producing classifications, which must be discussed in terms of the usefulness of the classification. In this thesis I will briefly examine some of the previously published methods for inferring phylogeny, and will suggest a unified approach to the estimation of phylogeny. It should be borne in mind that I am not discussing whether phylogeny ought to be used as the basis for classification.

Classical Phylogenetic Methods

Classical phylogenetic methods are somewhat ill defined. Three works which may be taken as basic sources for the explanation of these methods are Mayr, Linsley, and Usinger (1953), Simpson (1961), and Hennig (1966). Other major texts of taxonomy such as the books of Sokal and Sneath (1963), Davis and Heywood (1963), and Blackwelder (1967) do not adopt a phylogenetic approach to taxonomy and thus cannot be cited in this context. I will attempt to state some of the basic characteristics of classical phylogenetic methods as defined in these works, insofar as they define a single, coherent method.

The first characteristic is that groups are formed on the basis of the phenetic similarity of the organisms. Mayr et al. (1953) say, "One of the first steps in a phylogenetic study is usually the tabulation of the characters shared by the groups concerned." Simpson (1961) says,

> The observation and interpretation of characters in common do nevertheless play a large and essential role in evolutionary taxonomy, as they must in most systems of classification. They have other roles as well, but much of their importance is that they are one, and on the whole the chief, of the several criteria for judging propinquity of descent . . .

Hennig (1966) also bases his method on similarity of phenotype, but with specific restrictions on when this similarity is to be considered phylogenetically significant. This will be discussed later.

Phenotypic similarities are used only if it is felt that they result from common ancestry. There has been a great deal of discussion in taxonomy as to whether this practice is logically circular or not. If it were the claim of taxonomists to be detecting homology by use of knowledge of the phylogeny, the fears of circularity would be justified. But in fact taxonomists determine probable homology by degree of detailed similarity of the phenotypes of the characters involved. Simpson (1961) says,

> Homology does always involve characters in common, but it has also been sufficiently shown that the mere existence of characters in common or the possibility of abstracting an archetype or, its modern synonym, a morphotype is not a sufficient criterion of homology. . . . As far as characters in common are concerned, two criteria are fairly obvious: minuteness of resemblance and multiplicity of similarities.

This is based on the judgment of the probability that the similarities involved could arise twice independently.

> Intricate adaptive complexes are unlikely to arise twice in exactly the same way, hence to be convergent in two occurrences, and the probability

> of homology is greater the more complicated the adaptation and the closer the identity. On the other hand, similar adaptation with differences in characters not requisite for the adaptation as such is a strong indication of convergent homoplasy and opposed to homology [Simpson, 1961].

Another characteristic of classical phylogenetic methods is that they make use of information about which states of a character are primitive and which are derived. When two or more organisms share a derived state of a character, Hennig (1966) calls this "synapomorphy," the derived states being referred to as "apomorphous." He bases his entire method on synapomorphy: "It is evident that the presence of corresponding characters in two or more species is a basis for assuming that these species form a monophyletic group only if the characters are apomorphous, if their correspondence rests on apomorphy." However, he makes it clear that monophyletic groups are to be recognized not simply on the basis of sharing derived states, but on the basis of sharing steps in the "transformation series" of characters. For example, let species *A*, *B*, and *C* have a given character in states *a*, *a*', and *a*", where *a* is the most primitive and *a*" the most derived state, *a*' being intermediate. Hennig considers that this character indicates that *B* and *C* are a monophyletic group, since even though they are not morphologically similar, they share in their ancestry the step from *a* to *a*'.

The other authors do not make it clear exactly how information on primitive and derived states is to be used. Mayr *et al*. (1953) speak of using this information to determine the ages of groups: "When the primitive groups have been located and the primitive characters recognized, a rough approximation of the relative ages of the groups concerned is possible." Simpson (1961) discusses at some length the recognition of primitive states, but does not make it clear how this information is to be used in reconstructing the phylogeny. Hennig's is by far the best-defined system.

None of the works cited makes it clear what should be done in cases of incongruity. In any phylogeny, if two monophyletic groups overlap at all, one must be contained in the other. When groups are given which do not satisfy this condition, we have incongruity, in that it is impossible for all of the groups to be monophyletic. Hennig (1966) is the only author who faces the problem at all. He suggests that apparent incongruities may result from misinterpretation of the direction of change in some of the characters involved, from parallel evolution, or from incorrect evaluation of homology. If further examination of the organisms involved does not clear up the problem, he gives no concrete recommendation, saying only that "The common occurrence of parallelism and homoiologies, if not of pronounced convergences, indicates the necessity for phylogenetic systematics to take into account as many characters as possible in deciding kinship relations." A natural extension of Hennig's method would involve discarding those incongruous groups which share the fewest characters in common, or if there is no clear way to do this, discarding all incongruous groups.

In cases of incongruity, the authors cited do not treat all characters as of equal significance. Simpson (1961) says "Experience with particular groups always leads to empirically based judgment that some kinds of characters are more labile than others, and every specialist in classification acquires a 'feel' for the less labile or 'more reliable' characters in his group." Mayr et al. (1953) says "Taxonomic characters which are conservative (i.e., which evolve slowly) are most useful in the recognition of higher categories, those which change most rapidly, of the lower categories . . . "

Combining the principles stated above, we can outline a rough method, bearing greatest similarity to Hennig's methods, which should be a good approximation to classical phylogenetic methods. The steps are as follows:

1. Examine a number of characters in all of the organisms under study.

2. Eliminate those similarities likely to result from convergent evolution by using detailed resemblance as a criterion for probable homology.

3. For each character, decide which states are primitive and which derived.

4. Form all groups which share steps in the evolution of one or more characters. Computer programs to accomplish this step have been developed by Sharrock (1968).

5. If there is any incongruence (if any two groups overlap without either being contained in the other), resolve the incongruence by discarding those incongruent groups which share the least number of evolutionary steps in common. If the steps are to be considered as of different "weight," discard those groups sharing steps of least aggregate weight.

This procedure will result in a series of hierarchically nested sets of organisms, which are presumed to be monophyletic sets. From them an evolutionary tree can be drawn, although a time dimension is not specified and some parts of the tree may not have structure defined. The procedure is open to the objection that some of the information required, such as specification of primitive and derived states and of weights of evolutionary steps, may not often be available.

A more fundamental objection is that the procedure is not as well-defined as it appears to be. It is comparatively easy to construct data which contains incongruities which can only be resolved by following arbitrary rules with no biological justification. For example, suppose that we have five species, A, B, C, D, and E. Let group ABCD share five evolutionary steps. Let group BDE share four steps, and group BCDE share three. ABCD is incongruent with the latter two groups. If we wish to discard groups, we can either discard one well-demarcated group or two less well-demarcated ones. It is not immediately evident how this choice is to be made. One possible rule would be to discard all of the groups, so as to avoid arbitrary decisions. This has the disadvantage of discarding structure in the phylogeny produced. It is my experience that real data usually contains enough "noise" so that following this procedure one would discard all of the tentative structure of the phylogeny.

The Methods of Cain and Harrison, Wagner, and Throckmorton

A number of attempts have been made to produce well-defined and biologically well-justified procedures for inferring phylogenies. Each has difficulties which I will attempt to point out.

One of the first attempts was that of Cain and Harrison (1960). They recommend that enigmatic characters, in which it is not clear whether the basis of variation is genetic or environmental, be eliminated. All but one character in any group of characters which are functionally or ecologically necessarily related should be eliminated. For each pair of species a "distance" is then computed by summing the absolute values of the differences between the values of the characters in the two species, i.e., the distance between species _A_ and species _B_ would be $\frac{1}{n} \Sigma \left| X_i(A) - X_i(B) \right|$ where $X_i(A)$ is the value of character _i_ in species A. It is assumed that all characters are measured on quantitative scales, even if they are basically discrete characters. There are objections which can be raised to this measure of distance, which Cain and Harrison call the "mean character difference." It averages together numbers which have different units, and is highly sensitive to the units of measurement which the investigator happens to have chosen. But other measures of distance can be constructed which are not open to these objections and their substitution for the mean character difference is quite feasible. In calculating the mean character difference, Cain and Harrison recommend that for each pair of species, those characters whose similarity is thought to be due to convergence be dropped from the comparison.

They then recommend that standard phenetic grouping methods be used to make "patristic groups." Some of these phenetic methods are described by Sokal and Sneath (1963). The procedure is roughly as follows. Consider the species to be one-species "groups." Find the pair of groups which has the smallest distance, and combine them into a single new group, eliminating the original two. Find the distance from this new group to each of the other groups. This can be done in a number of ways. Three of the most common are calculating the distance between two groups as (1) the minimum species distance between them, (2) the maximum species distance between them, and (3) the average species distance between them. We now repeat the procedure and continue until all of the original species have been merged into one group. The method yields a hierarchy of non-overlapping groups. Cain and Harrison are not clear as to whether they want a hierarchy of groups or a series of disjoint, non-overlapping groups as their "patristic groups."

After the patristic groups are formed, they say that whatever cladistic considerations (i.e., relating to the form of the evolutionary tree) are available should be used to "regroup" the forms to obtain the probable phylogeny. While Cain and Harrison give a number of examples of such considerations, they make no statements of sufficient generality to establish a well-defined phylogenetic method. Important steps are ill-defined, and others are ill-justified, so that their methods can at best serve as a framework within which an investigator can construct his own method.

The procedure outlined by Wagner (1959) has many of the same sorts of problems. It is best described simply by quoting his description:

> To work out a phylogenetic problem three broad phases are involved: (a) systematic or comparative analysis of the plants in question to find and understand their contrasting characters; (b) determination of ground plans to find the character states common to all or most of the plants in order to deduce the most probable ancestral or primitive states; and (c) phylogenetic synthesis to assemble the taxa according to their respective deviations from the basic ground plan and from each other. The detailed steps are as follows: (1) to compare and study all the variable characters among the taxa; (2) to determine the generalized or primitive conditions on the principle that characters found in most or all of a number of related taxa are inherited essentially unchanged from the common ancestor, using data also from related taxonomic groups of the same level. (If no obvious trend can be determined in a given character that character may be used only for grouping purposes.) (3) to assign for each character the value 0 for the generalized or primitive condition, and 1 for the specialized or secondary condition (the intermediate states being assigned the value 0.5); (4) to list in tabular form the taxa and for each give the divergence values for the ground plan, both for individual characters and in total; and (5) to determine the mutual character groupings between taxa and then arrange them in sequence according to these groupings on a concentric chart or graph, the radii and branchings to be determined by the mutual character complexes, and the distances by the divergence indices. So that the facts may be made readily visual, the secondary or advanced states of each character should be expressed by letters (intermediate conditions, lower case; fully developed changes, upper case). Taxa are connected to each other by their ensembles of common features, which are plotted as the points of separation, i.e., as the most probable common ancestors. Such a method as this (although certainly subject to improvement and refinement) helps to solve problems. We can find correlations that have been overlooked. We are forced to use all available data and other workers can repeat our results with the same information. My method also shows at a glance the character groupings of the most probable common ancestors and thus outlines the pathways of phylogeny.

This is not really well-defined. In the critical step 5, it is not at all clear what is meant by "determine" and "arrange." From his description, it is possible to interpret the method as a system for the display of a phylogeny arrived at by unspecified methods. However, other authors such as Farris (1967) have interpreted the method as intended to produce phylogenies, and the next to last sentence of the above quotation seems to imply that the method is intended to do this.

A third method is that of Throckmorton (1962, 1965). The first step is the construction of "primary groups" by phenetic clustering methods, these groups being assumed to be monophyletic. Difficulties may be encountered when clearly-demarcated primary groups do not exist. Using the primary groups as units, we next construct a separate phylogeny of the group for each character, consistent with what is known about the direction of evolutionary change in the character. Such a phylogeny can be constructed whether or not the direction of change of the character is known, although it will be better constructed if direction is known. It is possible to some extent to infer the direction of change (Throckmorton, personal communication) but this complication will not be discussed here. The amount of variation within the primary groups can be used to make inferences about the gene pools of the ancestral populations. For example, if character state A is primitive and character state B is derived, and if we have three primary groups, one pure A, one pure B, and one containing some species with A and some with B, we can guess that the latter two groups had a common ancestor whose gene pool was segregating for the genes producing A and those producing B. From this ancestor two descendant species were produced, one with only B genes in its gene pool, and one heterozygous for the A genes and the B genes.

The phylogenies for the individual characters are now combined to produce an over-all phylogeny for the group. This step is not well-defined. In Throckmorton's 1965 paper, the example given for this step has little incongruity, so that serious difficulties do not arise. However, if each character indicates a different phylogeny, it is not clear how to combine the phylogenies. For example, if two-fifths of the characters indicate a phylogeny of one type, and three-fifths indicate a phylogeny of another type, should one accept the second phylogeny or some compromise between the two? In the examples worked in the 1962 paper, Throckmorton tends to solve the problem by indicating only that structure in the evolutionary tree which he feels is reasonably clearly indicated. For the reasons given above in the discussion of incongruence in classical methods, I feel that this procedure is not preferable.

The Method of Camin and Sokal

The first well-defined phylogenetic method to be suggested was the method of Camin and Sokal (1965). They assumed that the characters are coded into discrete states, and that for each character the sequence in which the states arise in evolution is specified, from primitive to most derived. Since these sequences have a tree-like form (in that more than one derived state can arise independently from the same primitive state), they will be referred to as character state trees. They also assume that there is no polymorphism in the ancestry of a group, so that each ancestor had exactly one state in each character. Finally, they assume that each state arises in evolution only from the state immediately preceding it on the character state tree. This rules out reversal of evolution, although independent origin of the same state in several different parts of an evolutionary tree are not ruled out.

Camin and Sokal's method is simply to find that evolutionary tree which requires the smallest number of evolutionary steps to explain the evolution of the group subject to the above assumptions. Fig. 1 illustrates the method for calculating the minimum number of evolutionary steps needed for a particular form of evolutionary tree. The bulk of Camin and Sokal's paper is devoted to proposing two methods for approximating to the solution in cases where there are too many alternative forms of tree for examination of each one to be feasible. Doolittle and Blombäck (1964) have also used this criterion in studies of amino acid sequences of a peptide of fibrinogen in five artiodactyls.. They try to find the phylogeny which requires the smallest number of amino acid residue changes in evolution.

Fig. 1.--Example of the method for calculating the number of evolutionary steps, given a phylogenetic tree.

Figure 1

ORGANISM	A	B	C	D
CHARACTER				
1	1	2	0	4
2	3	2	0	1

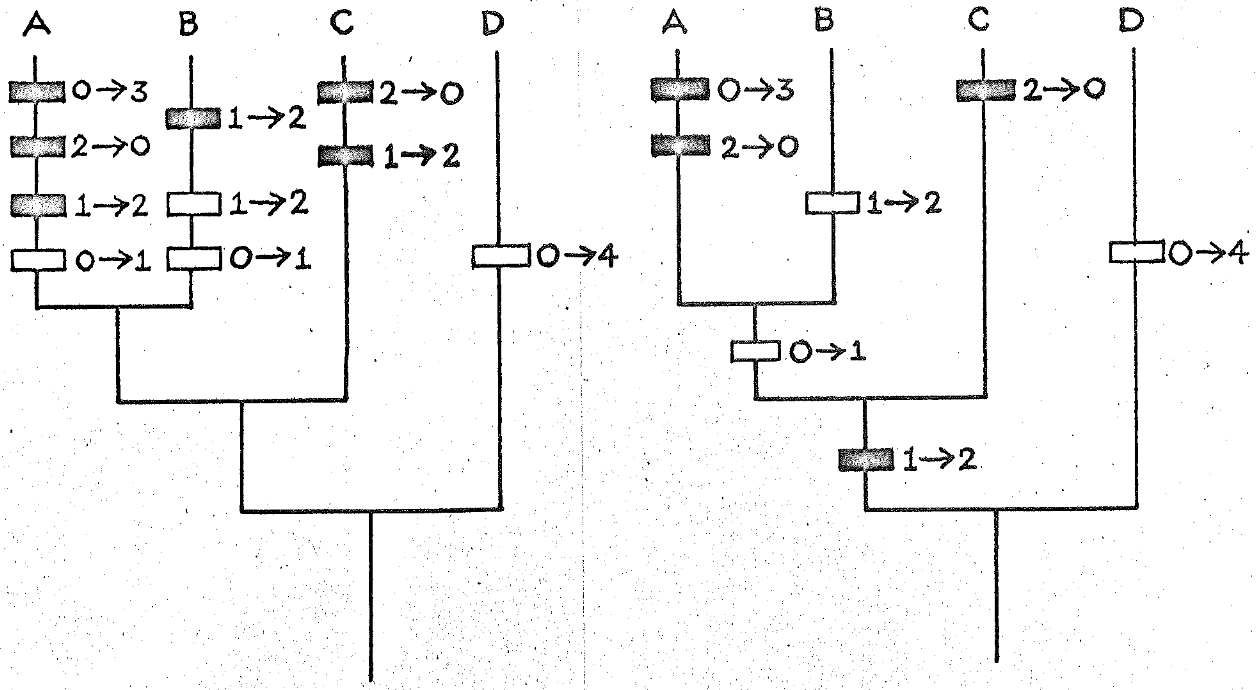
CHARACTER STATE TREES

CHARACTER 1 4 ← 0 → 1 → 2

CHARACTER 2 1 → 2 → 0 → 3

TREE BEFORE STEPS ARE MOVED

TREE AFTER STEPS ARE MOVED



Camin and Sokal refer to their method as "reconstructing cladistics by the principle of evolutionary parsimony." Parsimony implies an essentially aesthetic judgment of the economy of hypothesis necessary to explain observations. It is not clear that this can be equated with the criterion used by Camin and Sokal. It would seem that use of the term parsimony loads the dice in favor of any method to which it is applied. Camin and Sokal's method is better described as a minimum evolutionary steps method.

The Criticisms of Rogers, Fleming, and Estabrook

Rogers, Fleming, and Estabrook (1967) have made a number of criticisms of the method of Camin and Sokal. They point out that in any particular set of data, the evolutionary tree "most parsimonious" for one set of characters is not necessarily the same as the tree most parsimonious for another set, so that it cannot be maintained that the method will always produce the correct phylogeny. However, it is not clear that Camin and Sokal made such a strong claim. Rogers, Fleming, and Estabrook also point out that the most parsimonious phylogeny may require the existence of hypothetical organisms which have combinations of character states which are not viable. The criterion of minimum evolutionary steps thus contains an assumption of independence of different characters.

Their other two criticisms are fundamental. The first is that the probability of occurrence of the most parsimonious tree may be very small. They give the example that the probability that exactly fifty heads occur in one hundred tosses of a fair coin is only about 0.08. But this criticism misses the mark by viewing the situation prospectively rather than retrospectively. It is pointless to argue that the observed data have a low a priori probability of occurring, once it is known that they have in fact occurred.

They seem to be misreading the principle of "parsimony" as the statement that evolution will most likely happen in such a way that the minimum number of evolutionary steps will occur (and hence that most likely no evolutionary change at all will occur). This can be seen in the way that they pose the coin-tossing example: "As a matter of historical fact, a 'fair' coin has been tossed 100 times. Devise a scheme for guessing how many heads turned up." This question is simply irrelevant to the problem of inferring phylogeny. A better parallel to the type of problem encountered in guessing phylogeny would be the following: A coin tossed 100 times gives 50 heads. Devise a scheme for guessing the probability of heads. We guess 50 per cent, since in that case the observed results deviate least from their expectation. If the probability of heads is 50 per cent, there is an 8 per cent chance of getting 50 heads in 100 toses, while when the probability of heads is (say) 70 per cent, there is only a 1.3 per cent chance of getting 50 heads in 100 tosses. Thus, even though the chance of getting 50 heads is low a priori in all cases, an a posteriori view shows that we are more likely to get 50 heads if the probability of heads is 50 per cent than if it is 70 per cent. This is the method of maximum likelihood, of which more below.

The principle of "parsimony" or of minimum evolution states that given that the observed data have occurred, it is most likely that they occurred according to whichever phylogeny would require the smallest amount of "evolution." Though this statement may be on shakier ground than the statement that the minimum possible number of evolutionary steps will occur, it is certainly more useful in selecting a phylogeny.

The remaining criticism of Rogers, Fleming, and Estabrook is that "the most likely step at any given point in time may not be a step whose result is a most parsimonious tree for the objects that were eventually evolved."

This criticism is not substantially different from the preceding one, and will not be discussed in detail. It evolves the same a priori point of view.

The Maximum Likelihood Approach

Despite the weakness of their major criticism, Rogers, Fleming, and Estabrook take an important step in discussing "parsimony" in terms of probability. This changes the terms of discussion from the ill-defined concepts of "parsimony" and amount of evolution to the more clearly defined concept of probability. Substituting probability for "parsimony" leads to the statement that the most likely phylogeny is the one on which the observed data have the highest probability of arising. This is identical with the statistical method of maximum likelihood, provided that we view phylogeny as the parameter being estimated.

The use of a conceptual framework involving probabilities was foreshadowed by a comment of Camin and Sokal (1965) that "the method as described above assumes equal probability of all evolutionary steps after the characters have been coded." They attribute this insight to E. C. Minkoff. In fact, the method of maximum likelihood had already been used by Edwards and Cavalli-Sforza (1964). They worked on the frequencies of alleles in blood group polymorphisms in human populations, attempting to derive a phylogeny. The first step of their analysis is to transform the allele frequencies onto a set of rectangular co-ordinates. The transformation was chosen so that the process of random genetic drift with multiple alleles would be transformed into a process of Brownian motion in many dimensions. Given an evolutionary tree, each fork of which has associated with it a time and a set of allelic co-ordinates, it is possible to calculate for each segment of the tree the probability that the stated amount of change would occur during the stated time interval. They calculate the likelihood of the whole tree as the product of the likelihoods of its segments, since Brownian motion during successive time intervals is independent. They have developed methods by which a good guess can be made at the maximum likelihood tree without examining all possible forms of tree. The phylogenies obtained in this way agree surprisingly well with standard interpretations of the evolutionary origin of human populations (Cavalli-Sforza and Edwards, 1965). It is interesting to note that before developing their maximum likelihood method, Edwards and Cavalli-Sforza used a "method of minimum evolution." The results from this method were not identical to those from the maximum likelihood method, although they were very similar.

The remainder of this thesis will expand on the method of maximum likelihood as applied to the estimation of phylogeny, and apply it to various types of data which are commonly encountered. Hopefully, application of a statistical inference approach to the problem will ultimately lead to the elaboration of phylogenetic methods which are well-defined, biologically well-justified, and of considerable power. It cannot be claimed that we are anywhere near this goal at present.

GENERAL PRINCIPLES

Types of Evolutionary Trees

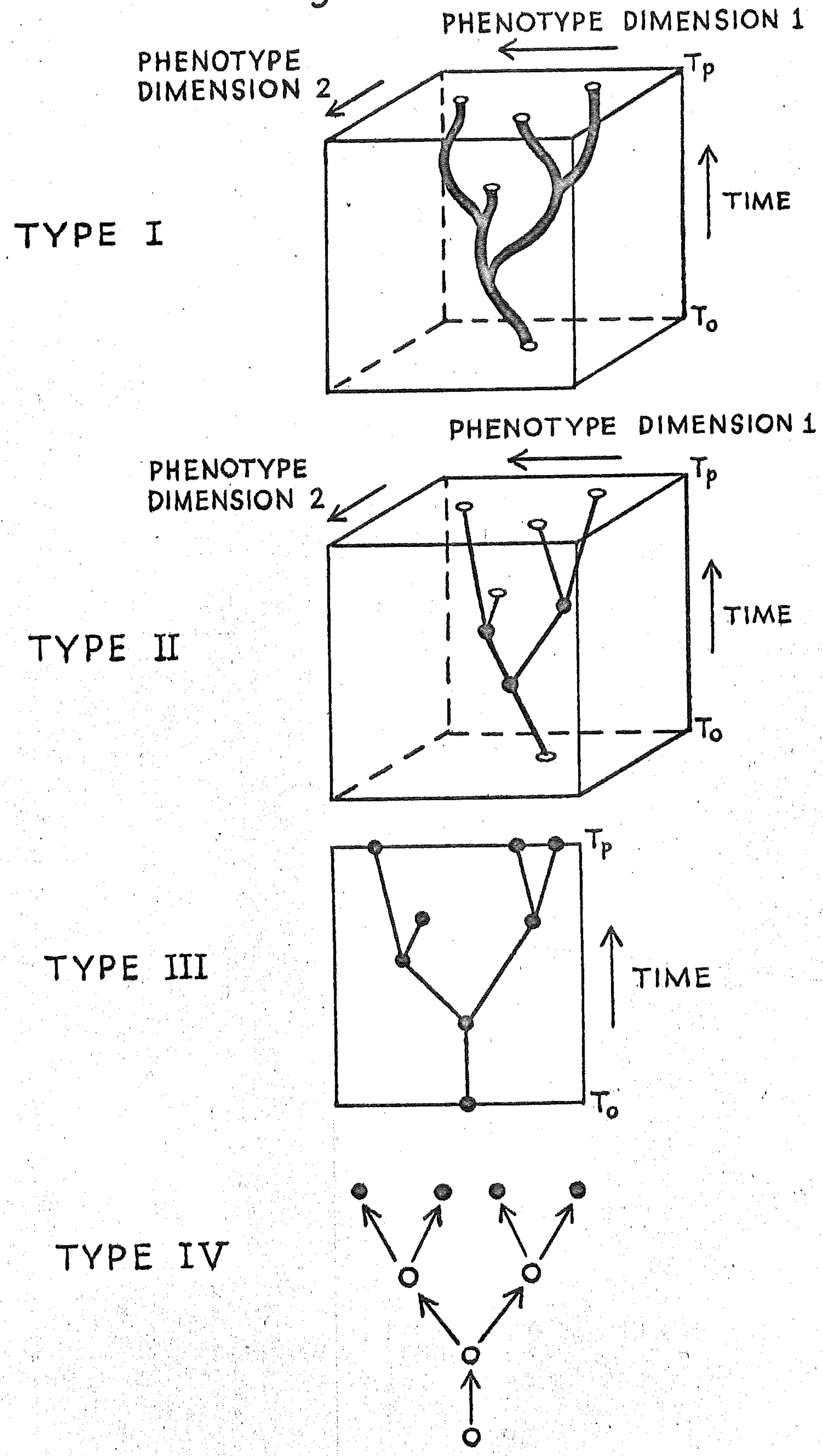
The terms "evolutionary tree" and "phylogeny" have been used above without definition. When used in the literature, these terms have no precise meaning, denoting a variety of related but distinct concepts. In this section, I will define four types of evolutionary tree.

A representation of the total evolutionary history of a group would include a pedigree showing every individual that ever existed. For each individual the genotype, phenotype at each moment, and trajectory in time and space would be given. Present phylogenetic studies work far below this level of precision. No attempt is made to infer the phenotypes of past individuals, other than those observed as fossils. At best, an attempt is made to infer the phenotype parameters of past populations, and to draw a pedigree of populations. Since speciation is represented on such pedigrees as an all-or-none, instantaneous process, the pedigrees usually have the shape of a branching tree. While such a representation of evolution is obviously crude and oversimplified in the extreme, it is probably the best that can be attempted with the data available.

Four types of evolutionary tree are depicted in Figure 2. Type I gives the most complete description of the evolution of the group in question. There is a finite number (in this case four) of populations at the tips of branches. At time T_o there is only one population ancestral to these tip populations. Every population ancestral to the populations at the tips of the

Fig. 2.--Four types of evolutionary trees.

Figure 2


PHENOTYPE DIMENSION 1
PHENOTYPE DIMENSION 2
Tp
TIME
To
TYPE I
PHENOTYPE DIMENSION 1
PHENOTYPE DIMENSION 2
Tp
TIME
To
TYPE II
Tp
TIME
To
TYPE III
TYPE IV

branches and descended from the population at time T_o is represented. Each population is described in terms of a series of phenotypic parameters (in this case, two). With a sufficiently large number of phenotype parameters, this type of tree can convey an almost complete record of the evolutionary history of a group.

The tree of type II is derived from the tree of type I by omitting information about all populations except those at the tips of branches, the ancestral population at time T_o, and the "latest common ancestor" populations at the forks of the tree. These populations have their times and phenotype parameters specified, as well as their ancestor-descendant relationships, symbolized in the figure by thin lines connecting them. Further removal of information results in a tree of type III, in which the times and ancestor-descendant relationships of the populations are given, but not phenotype parameters. The horizontal dimension of the diagram is formally meaningless, serving only to separate the populations. Finally, omitting the times of occurrence of the latest common ancestors brings us to a tree of type IV, in which only the order of branching is shown.

It is possible to make rigorous definitions of mathematical entities similar to the trees described here, but there seems to be no point in doing so in this thesis. It is also possible to define other kinds of trees. One further type of tree which often occurs in the literature is the type which would result if information on the phenotype parameters, but not the times, of the latest common ancestors were given. In the literature of phylogenetic taxonomy, the names used for different types of trees are extremely inconsistent. It would be desirable to have a standardized nomenclature for evolutionary trees. In its absence, I will refer to the types of trees by their numbers.

The Application of Bayesian Inference and Maximum Likelihood

The construction of a phylogeny is the process of making inferences about the evolutionary history of a group of organisms. It is very similar to processes of statistical estimation. In doing phylogenetic inference, we are given information about the phenotypes of a group of organisms and assumptions about the evolutionary processes responsible for the diversity of the group. We are required to choose among a set of possible evolutionary trees for the group. In a typical problem of statistical estimation, we might be given a series of numbers and the assumption that they were drawn at random from a Normal distribution with unknown mean and variance. We are required to choose values of the mean and variance from the set of possible values.

The central theme of this thesis is the feasibility and usefulness of considering phylogeny as a problem in statistical inference. A statistical approach is no more free from arbitrary assumptions than are classical methods. But it does have the advantage of making clear the extent and manner of dependence of the specific procedures followed in making the inference on the biological model assumed. The possibility of examining this dependence comes from the fact that once a general method of estimation is chosen, for every biological model of evolution which we assume, we can generate an estimation procedure for inference of the evolutionary tree. We can even attempt to find biological models whose statistical estimation procedure is the same as a classical phylogenetic method, although success is not guaranteed.

Even if statistical inference methods are not adopted, a statistical vocabulary should be used in phylogenetic studies. Many of the classical methods are stated in language which implies that with their use the evolutionary tree can be infallibly deduced. Elsewhere in the same works, the authors often clearly indicate that they are well aware of the uncertain and tentative nature of the results obtained by use of these methods. If only for self-consistency, phylogenetic methods should be described in terms of probability and likelihood. A misleading aura of infallibility can scarcely be desirable.

It is not my intention to review philosophies of statistical estimation. A variety of approaches exists, each yielding estimators with particular desirable properties. The variables being estimated in a phylogenetic problem include not only continuous variables such as times of branching of the tree, but also discrete ones such as the topological shape of the tree. Many of the existing approaches to estimation cannot be used for this reason, since properties such as lack of bias, minimum mean square error, and efficiency are meaningless when applied to the discrete variables. I will make use of the related Bayesian inference and maximum likelihood methods, because they are almost always applicable, have a simple interpretation, and have various desirable properties in those situations in which the properties are meaningful.

Bayesian inference makes use of Bayes' Theorem. Suppose that we have a series of events $A_1, A_2, \ldots, A_n$, corresponding to forms of the evolutionary tree, one of which must be correct. Suppose that we have *a priori* probabilities $P(A_i)$ that evolution occurs according to hypothesis *i*. Suppose that we have a series of possible configurations or outcomes of the observed data: $B_1, B_2, \ldots, B_m$. We are given the probability that data outcome B_j occurs, given that hypothesis A_i about the tree is correct, and this probability is denoted by $P(B_j|A_i)$. By the standard definition of conditional probability we have $P(A_iB_j) = P(A_i)P(B_j|A_i)$, where $P(A_iB_j)$ is the probability that evolution occurs according to hypothesis *i* and the data outcome is B_j.

We want to find the probability that, given that data outcome B_j occurs, the evolutionary tree was of type A_i. This probability is written $P(A_i|B_j)$ and is equal to $P(A_iB_j)/P(B_j)$ by the definition of conditional probability. We have not been given $P(B_j)$, but it is equal to $P(A_1B_j) + P(A_2B_j) + \ldots + P(A_nB_j)$ so that we can write

$$P(A_i|B_j) = \frac{P(A_i)P(B_j|A_i)}{\sum\limits_{\text{all } k} P(A_k)P(B_j|A_k)} \qquad (1)$$

which is Bayes' Theorem, found in any text of probability.

If we are given a priori probabilities $P(A_i)$ of the hypotheses, and conditional probabilities $P(B_j|A_i)$ of the data given the hypotheses, we can use Bayes' Theorem to find a posteriori probabilities of the hypotheses given that data of type j has actually been observed. If we want to choose one of the hypotheses as our estimate, it is natural to choose the one with the highest a posteriori probability. Since all the expressions for $P(A_i|B_j)$ which have the same B_j have the same denominator in equation (1), this amounts to choosing the A_i which has the maximum value of $P(A_i)P(B_j|A_i)$, which is $P(A_iB_j)$. If the a priori probabilities $P(A_i)$ are known to be or assumed to be equal, this will be equivalent to choosing that value of A_i which gives the maximum value of $P(B_j|A_i)$. This procedure is often followed when the a priori probabilities are unknown. $P(B_j|A_i)$ is called the likelihood of hypothesis A_i given data B_j, and the procedure is known as the method of maximum likelihood.

Bayes' Theorem is often written in the odds form:

$$\frac{P(A_i|B_j)}{P(A_k|B_j)} = \frac{P(A_i)}{P(A_k)} \cdot \frac{P(B_j|A_i)}{P(B_j|A_k)}$$

Thus, given the data outcome B_j, the odds favoring A_i over A_k are the product of the a priori odds and the likelihood ratio of A_i versus A_k. Obviously only the A_i with the greatest value of $P(A_i|B_j)$ will have this ratio greater than unity for all A_k.

Bayesian and maximum likelihood methods have a number of desirable properties. It is easily shown that a Bayesian method is the method of estimation which has the highest probability of being correct, as follows: Given data outcome B_j we must decide which of the A_i to estimate. If we choose A_k, our probability of being correct is $P(A_k|B_j)$. To minimize the over-all probability of error, we must minimize the probability of error for each $\underline{j}$. This can be done by choosing the $\underline{k}$ which gives the highest value of $P(A_k|B_j)$. But this is exactly the procedure we follow when we use a Bayesian method to get a point estimate of the evolutionary tree (i.e., to guess a single tree). This property is less meaningful when we have a continuum of possible trees, each with a probability of zero. It should also be noted that the property of having minimum probability of error does not apply to maximum likelihood methods. In them we choose the A_i with the highest value of $P(B_j|A_i)$, which is not the same as choosing the A_i with the highest value of $P(A_i)P(B_j|A_i)$, especially if the a priori probabilities differ substantially. It is easy to construct cases in which $P(A_i)$ is zero, but $P(B_j|A_i)$ is larger than any other $P(B_j|A_k)$. In these cases, maximum likelihood methods will choose tree A_i, while Bayesian methods will not.

Both Bayesian and maximum likelihood methods have the property of being consistent. This means that as more and more independent characters are used to make the estimate, the probability that the wrong tree will be chosen will approach zero, so that with sufficiently many independent characters we can reach an arbitrarily high level of accuracy. In such a situation $B_1, \ . \ . \ . \ , B_m$ represent the outcomes of a single character. The data outcome can be represented by an n-tuple $(n_1, \ . \ . \ . \ , n_m)$, where n_j is the number of characters in which the outcome is B_j. Bayes' Theorem can be stated in odds form:

$$R_{i/k} = \frac{P[A_i|(n_1, \ . \ . \ . \ , n_m)]}{P[A_k|(n_1, \ . \ . \ . \ , n_m)]} = \frac{P(A_i)P[(n_1, \ . \ . \ . \ , n_m)|A_i]}{P(A_k)P[(n_1, \ . \ . \ . \ , n_m)|A_k]}$$

Kendall and Stuart (1961, vol. 2, p. 40) have presented a general (although not elementary) proof that as one increases $N = n_1 + n_2 + \ . \ . \ . + n_m$, the probability that the likelihood ratio

$$\frac{P[(n_1, \ . \ . \ . \ , n_m)|A_i]}{P[(n_1, \ . \ . \ . \ , n_m)|A_k]}$$

is greater than unity approaches one. This proves that the maximum likelihood method is consistent. Since a Bayesian method with correct values of the $P(A_j)$ has a smaller probability of error than the corresponding maximum likelihood method, Bayesian inference is also consistent. It can even be shown for simple cases that even when the wrong values of the $P(A_j)$ are used, provided that $P(A_i) > 0$ for the true value A_i, the estimation is consistent.

It should be kept in mind that the proofs of consistency depend on the different characters behaving as if they were identical, i.e., $P(B_j|A_i)$ must be the same for all characters. This requirement is not met in most real data, so that, strictly speaking, we cannot be sure that consistency holds in any real case.

In addition to point estimation, we can do interval estimation, in which the estimate is a set of trees rather than a single tree. A simple procedure which suggests itself in the case of Bayesian inference is to construct the set, given a data outcome B_j, by first placing in it the A_i which has the largest value of $P(A_i|B_j)$, then the next largest, and so on until a predetermined amount a of probability has accumulated. The resulting set will contain the correct tree a of the time. We cannot, however, call a the confidence level of the set, since confidence intervals are constructed differently. I will not deal further with the question of interval estimation. The validity of all estimation procedures is dependent on the validity of the models of evolution on which they are based. Since the models which are usable at present are so crude that they can only be termed gross, it is likely that one deludes oneself by making statements that a given set has a 95 per cent chance of containing the correct evolutionary tree. It might be better to stick to point estimation, and simply hope that the true evolutionary tree looks somewhat like the estimated one.

Probability Models of Evolution

In all of the above procedures and considerations, the probabilities $P(A_i)$ and $P(B_j|A_i)$ play a crucial part. The values of these probabilities are determined by the model of evolution which is assumed. It should be remembered that $P(B_j|A_i)$ is the probability that the data outcome will be B_j given that the true tree is A_i. The interpretation of this quantity as a probability need not imply that the mechanism of evolution is random or probabilistic in any way. If the entire process is deterministic, and if we know the mechanisms, $P(B_j|A_i)$ will be either 0 or 1, depending on whether evolution according to tree A_i results in data of type B_j. In this case, since Bayesian inference maximizes $P(A_i)P(B_j|A_i)$, the estimate will be the tree which gives rise to data of type B_j and which has the highest *a priori* probability $P(A_i)$. Maximum likelihood estimation, since it is equivalent to

Bayesian inference with the $P(A_i)$ equal, will not choose among those A_i which have $P(B_j|A_i)$ equal to one, if several such exist.

Thus the use of probabilities does not necessarily imply acceptance of a probabilistic model of evolution, which may be unacceptable to many biologists. Since we are dealing with events which are essentially unique (it being impossible to run repeated trials of the evolution of the vertebrates) we must think of the probabilities as measuring our own uncertainty as to what happened in evolution, rather than describing any natural random process. Even when evolution is completely deterministic, we can have apparent randomness arising from uncertainty about the initial conditions of the group before the relevant portion of its evolution. And it should not be overlooked that randomness in the data can arise from processes of sampling, either human or natural (as with fossilization), which give rise to the particular data under consideration.

So far, I have assumed that the $P(B_j|A_i)$ are available. Sometimes, however, we may be uncertain as to details of the models of evolution itself. Thus, we may have three possible models, M_1, M_2, and M_3. If we have <u>a priori</u> probabilities of the models, reflecting our intensities of belief in them, we can write

$$P(B_j|A_i) = P(M_1)P(B_j|A_i, M_1) + P(M_2)P(B_j|A_i,M_2) + P(M_3)P(B_j|A_i,M_3)$$

where $P(B_j|A_i,M_k)$ is the probability that the data outcome is B_j given that the model is M_k and the true tree is A_i. Once $P(B_j|A_i)$ is calculated, the estimation can proceed in a normal manner. We may also wish to make an estimate of the model of evolution as well as of the evolutionary tree. In this case we are estimating the pair (A_i,M_k) where A_i is the tree and M_k is the model. Bayesian inference chooses that pair which maximizes

$$P(A_i,B_j,M_k) = P(M_k)P(A_i|M_k)P(B_j|A_i,M_k).$$

In many cases, $P(A_i|M_k) = P(A_i)$ when the _a priori_ probability of the form of evolutionary tree is independent of the model. If either or both of the _a priori_ probabilities $P(M_k)$ and $P(A_i|M_k)$ is not known, we can assume that all of the alternatives are equiprobable _a priori_, in which case we are carrying out either maximum likelihood estimation or a mixture of Bayesian inference and maximum likelihood estimation.

When we estimate both the tree and the model, we inevitably lose accuracy in each one individually. The more parameters are being estimated at once, the closer the data will appear to fit the model. For example, if we leave the model of evolution completely unspecified, and pick an evolutionary tree at random, say A_1, we can always invent a completely artificial and contrived model of evolution under which $P(B_j|A_1) = 1$ while $P(B_j|A_i) = 0$ for all other A_i. Under this model, the precision of the estimation of A appears to be total, when in fact it is nil, for we could as easily have chosen any other tree A_i. In general, accuracy of estimation of the evolutionary tree requires that we estimate as little as possible of the details of the model of evolution from the same data. If we do not have _a priori_ probabilities for the models M_i, but are not particularly interested in estimating anything but the tree, the best that we can do is to treat the _a priori_ probabilities of the models as equal. This is equivalent to doing maximum likelihood estimation of the model and then discarding the resulting estimate. The fact that we ignore information about M does not make the information about A more accurate. It is important to be able to recognize this type of "silent estimation" so as to be able to assess its effect on the accuracy of the tree estimation.

Models for Generating the Form of Trees

When we estimate trees of type I or type II (see Fig. 2), it at first seems that the nature of these types of trees forces us to assume a deterministic model of evolution. Both of these types of trees specify the phenotypes of the ancestor species at the forks of the tree and the phenotypes of the species at the tips of the branches. The specifications of the tree then include the data. Thus the probability of the data outcome given the tree must be either 0 or 1 if there is no sampling error in the data. The probabilistic model of evolution enters into determining the a priori probability of a tree.

It is useful in discussing this type of problem to separate conceptually the evolution of the form of the tree from the evolution of the phenotypes of the organisms on it. Let T represent the information about the times of occurrence of the tree root, the latest common ancestors, and the branch tips. Let Q represent the information on the phenotypes of the populations represented by the tree. In Bayesian inference of a tree of type I or II, we maximize over all (T, Q) the quantity

$$P(B_j,T,Q) = P(B_j|Q)P(Q|T)P(T).$$

The difference between estimation of type I trees and type II trees consists of whether or not Q contains specifications for the phenotypes of populations not at the root, at forks, or at branch tips. The quantity $P(B_j|Q)$ is either 0 or 1, depending on whether Q includes the data configuration B_j in its specification of the phenotypes at the branch tips. In cases with sampling error, where B_j consists of sample parameters and Q of population parameters, this is not true. In trees of type III and IV, Q consists only of information about the branch tips, and since no aspect of Q is contained in the specifications for the tree, the maximization is of the quantity

$$P(B_j,T) = P(B_j|T)P(T)$$

over all possible values of the parameters comprising T. We usually have a model specifying $P(Q|T)$, but we do not often have a model specifying $P(T)$. When we have such a model, we can do Bayesian inference, but when we do not, we are forced to assume the $P(T_i)$ all equal, so that we are doing a mixture of Bayesian inference and maximum likelihood. In the case of trees of type I or II, we are in effect finding that tree of all those which could lead to data of type B_j which has the highest probability a priori.

Discussion in any detail of the model of change of the phenotypes must be deferred to later sections, but it is appropriate at this point to discuss models of evolution of the tree form, as given by the parameters comprising T. In reality it is impossible to separate the processes of phenotypic change from those of speciation. If the characters under consideration are ones which contribute to or reflect the species isolation, change in the characters would be expected to be correlated with speciation. However, for a first approximation it will be assumed that the processes are independent. Models incorporating the dependence would be desirable, but appear difficult to state at present. As was noted earlier, it is already a crude approximation to consider speciation as an all-or-none, instantaneous process. One simple model for the branching of evolutionary trees is the Yule process (Cavalli-Sforza and Edwards, 1967), a classical model in probability theory (see, for example, Feller, 1957). In such a process during a time interval of length dt, each line has an independent probability $s\,dt$ of splitting (for very small values of dt). It is possible to derive expressions for the probabilities of generation of different trees, but the matter is fraught with difficulties. The paper of Cavalli-Sforza and Edwards contains some comments on the problem. Since there will be no need for the probabilities here, they will not be derived.

In estimation of trees of types I, II, and III, we are required to estimate the times of splitting of the ancestral lines. If we are willing to accept a particular model of generation of the tree, we can use it to obtain P(T) and do Bayesian inference. Type IV trees present a more severe problem. If we have a model of tree generation, we can work out the probability of a particular topological form of tree by summing the probabilities of the trees of type III consistent with the particular type IV tree over all possible positions in time of the fork points. If we do not have a model of tree generation, we must do maximum likelihood estimation, as mentioned above. Since the positions in time of the fork points will affect the probabilities of phenotypic change $P(Q|T)$, when we estimate tree form by maximum likelihood we must estimate these times. If we are to end up with a tree of type IV, we must discard these times. This is an example of "silent estimation," which was mentioned above.

The models of tree form generation should, of course, take extinction into account. The Yule process does not. It is easy enough to state simple models involving extinction, but it is quite another matter to derive usable expressions for the probabilities of different tree forms. This is the usual quandary. In addition to extinction, a moderately realistic model of tree form generation would take into account the sampling of species for inclusion in the study. If the tree is generated by a Yule process, and the species to be included in a particular study are sampled from the tree in some random manner, the tree form of the selected species would be dependent not only on the process which generated them, but also on the type of sampling used. If we cannot say that the species in a given study were sampled in some random or "unbiased" way from the relevant group, we may be forced to use a maximum likelihood approach even when we have some idea of the process generating the tree for the larger group. Similar difficulties arise when we use taxa higher than the species as our basic units. We can use "generic characters" and assume that they characterized a single species ancestral to the genus, but it is likely that we will become ensnared in problems of estimating the times of occurrence of these ancestral species, which will in effect be the tip species on the tree.

The Treatment of Fossils

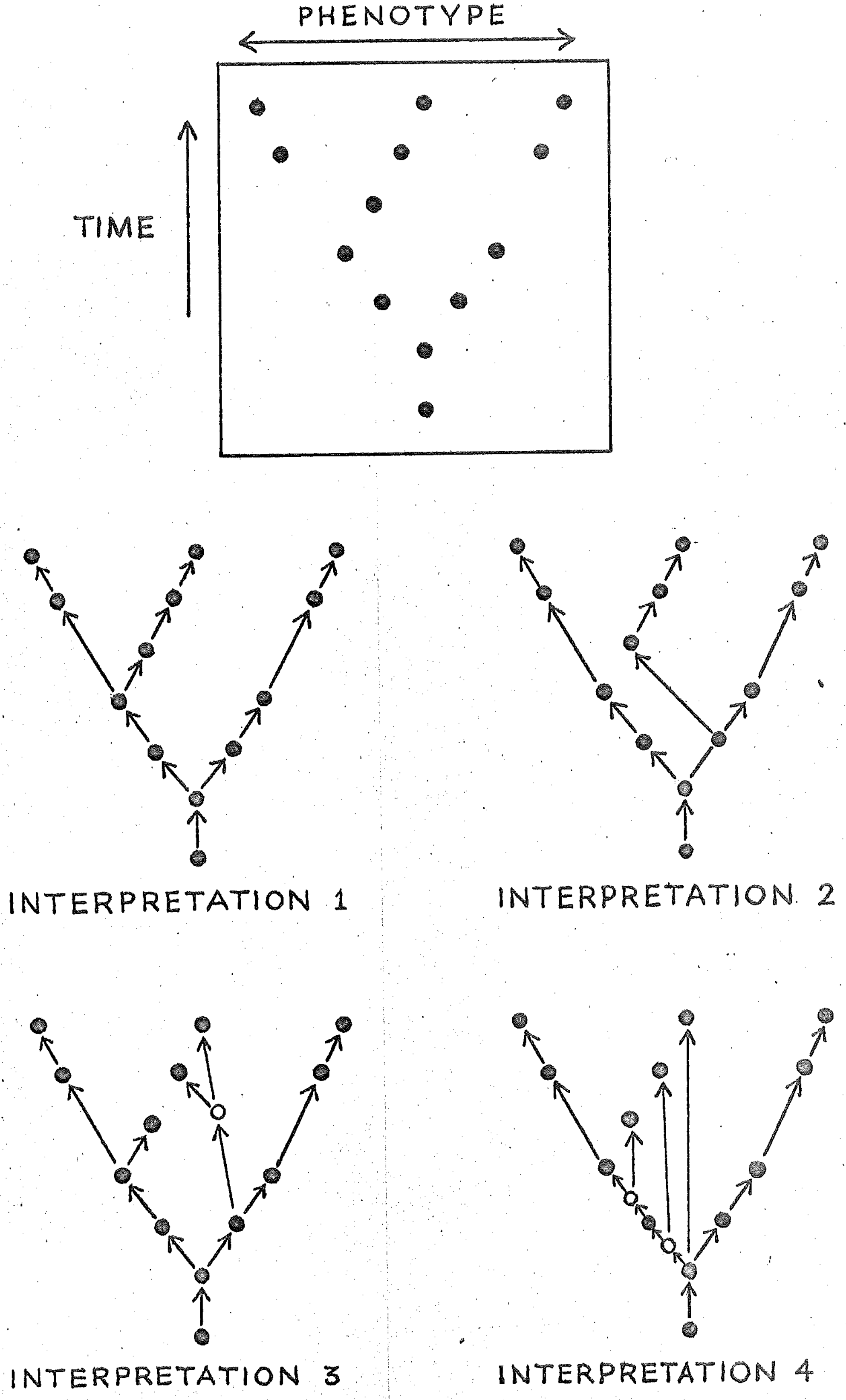
If we look at problems of estimation of tree form and fork times in the above framework, we may be led to some clarification of the role of fossil evidence in inferring evolutionary trees. It is commonly held to be of crucial importance. Sokal and Sneath (1963) say that ". . . since we have only an infinitesimal portion of phylogenetic history in the fossil record, it is almost impossible to establish natural taxa on a phylogenetic basis." Hennig (1966) feels that it is quite possible to obtain phylogenies without fossils. However, he assigns special significance to them in assigning times of origin to groups: ". . . fossils with relatively apomorphous characters can be very important . . . because they not only prove the existence of the groups to which they belong but also may prove the simultaneous existence of other groups with strongly plesiomorphous characters." Both classical methods, which often use fossils to infer phylogeny, and Hennig's methods, which use fossils to assign times to the origin of groups, contain the assumption that the fossils represent the ancestors of the modern groups which they resemble. Hennig (1966) makes the assumption explicit by saying "all this is true, of course, only if convergence or retrogressive development can be excluded."

The straightforward treatment of fossils is to treat them as additional species in the analysis, according them no special treatment aside from taking into account their time of occurrence. This does not bias us in any way from discovering convergence in fossils, as other treatments might. If we look at the alternative interpretations in Fig. 3, we can see that we might favor interpretation 1 over interpretation 2 because it requires less drastic change in any phyletic line, and hence is presumably more likely. Interpretation 1 is favored over interpretation 3 because it requires only two speciation events rather than four, and we may consider speciation improbable relative to phenotypic change. If for some reason we felt that speciation was far more probable than phenotypic change, we would prefer interpretation 4 to interpretation 1.

Classical methods which assign ancestral positions to fossil organisms implicitly assume that speciation is improbable compared with phenotypic change of the order of magnitude observed. This sort of judgment is involved, for example, when _Australopithecus_ and _Homo erectus_ are treated as ancestors of _Homo sapiens_, rather than as organisms sharing common ancestors with _sapiens_ but not ancestral to it. The interpretation of fossils as ancestors does not depend on special treatment of fossils in the analysis, provided the proper assumptions are made about the probability of speciation. The arbitrary nature of according special treatment to fossils can be seen if we consider a series of hypothetical organisms, one collected in 1967, one collected in 1867 and stored since then, one fossilized in 1967 B.C., and one fossilized in 1,000,000 B.C. Which should be given special treatment?

Fig. 3.--Alternative interpretations of a hypothetical example in which both fossil and recent organisms are included.

Figure 3



INTERPRETATION 1

INTERPRETATION 2

INTERPRETATION 3

INTERPRETATION 4

A statistical viewpoint gives no support to any special treatment, provided that the time at which the organisms existed is taken into account.

Dependence of Characters

A statistical approach can also be of value in considering the effect of dependence of characters in taxonomic practice. There are several levels at which characters may be dependent on one another. Characters in the same individual may be physiologically or genetically interdependent. This affects the probability of simultaneous evolutionary change in the characters. Suppose, for example, that a character U has probability p of changing from state U_1 to state U_2 in a particular portion of an evolutionary tree. Suppose that character V also has probability p of changing from state V_1 to state V_2 in the same portion of the tree, for the simple reason that U and V measure the same property of the organisms in different ways. Then the probability that both changes occur together in the same portion of the evolutionary tree is not p^2 but p. If characters which are correlated in this way are treated as if independent, there will result misevaluation of the probabilities of evolutionary trees. Any genetic or physiological correlation must be built into the probability model. Alternatively, if we are forced to assume independence of characters, we can recode the data to take the known dependences into account. This provides a justification of the procedure of Cain and Harrison (1960), who urge that of any group of necessarily correlated characters, all but one be discarded. Of course, if the dependence can be built into the model, there will be no loss of information, so that this is to be preferred to discarding characters.

There can also be non-independence of characters between individuals of the same species. Examples of this are polymorphism and sexual dimorphism.

In a simple one-gene polymorphism with three phenotypes, the occurrence of the heterozygous phenotype is obviously dependent on the occurrence of both homozygous phenotypes. If we have three species, in one of which we have four specimens, two of type AA and two of type aa, in another of which we have two specimens of type Aa, and in the third of which we have ten specimens of type aa, it is clear that the first species is more similar to the second than either is to the last. This might superficially seem paradoxical, since in the first and third we have specimens with the same phenotype, while in the first and second we do not. The resolution of the paradox is, of course, that our basic units in studies of higher categories are populations rather than individuals. The allele frequencies are the characters of the population in cases of polymorphism (or the morph frequencies if the genetic basis of the polymorphism is unknown).

Hennig (1966) has pointed out that, as we do not often know the genetic basis of variability or the position of species boundaries, detection of polymorphism and resolution of intraspecific differences ". . . can be regarded as a systematic problem of the 'lowest taxonomic units,' which we considered to be the individuals." Two approaches to this problem are possible. The more general and internally consistent approach is to make these judgments part of the phylogenetic estimation procedure. Specimens would be aggregated into species by the procedure in order to reduce the number of speciation events which it would be necessary to assume, at the cost of assuming that the species were polymorphic. The alternative method would be to make decisions about species boundaries prior to using statistical inference. The judgments as to which differences are specific and which polymorphic are then built into the probability model. If these judgments are correct, the inclusion of this information in the model will improve the resolving power of the phylogenetic inference above the species level. If there is no model giving probabilities of speciation events, the first of these two approaches cannot be carried out. The second also allows use of gestalt information not formally coded in the data.

Sexual dimorphism has many of the same problems as polymorphism. In particular, if we have a series of males and females which differ, we must decide whether the difference is sexual dimorphism or species difference. The possible approaches are the same as given above. Sexual dimorphism is a type of character dependence. Although it may superficially appear to be equivalent to polymorphism, it is quite different in that it does not demand maintenance in the population of any genetic heterogeneity other than for the genes determining sex. If it is represented as polymorphism, the probabilities of evolutionary change will be misleading, so that this approach is quite unsatisfactory. On the other hand, we cannot code a sexually dimorphic character as two separate characters unless the probability model of evolution takes into account the dependence of the two characters. If the male form of a character changes, this may well be correlated with a change in the female form. We get into the same sorts of problems discussed above with regard to the probability of simultaneous change in the two characters. If the probability model assumes independence of evolution in different characters, we must code a sexually dimorphis character as a single character, the "state" of which consists of the states of the male and female forms. Change in either or both will result in a new "state" of the compound character.

The third level of dependence is dependence of characters in different species. Under this heading can be placed effects due to competition, predation, parasitism, and mimicry. In effects of this type, the change of a character in one species will have an influence on the probability of change of a character in another species. The models presently in use in numerical phylogenetic methods do not take these phenomena into account. Since members of a closely related group of species do not often prey upon, parasitize, or mimic each other, these three phenomena can often be treated as if they were environmental effects. However, closely related species often compete with each other. The characters chosen in taxonomic studies frequently reflect feeding and locomotor activities and size, aspects of the organism likely to be strongly influenced by competition. Treating the evolution of different lines as independent implicitly assumes the absence of competition. This is a major difficulty with present models which will not be tackled in this thesis because of the complexity of the problem.

Dependence can result from two characters being affected by the same selection pressure. For example, if a particular adaptation can be made by altering either character X or character Y, then the probability of change in character Y (which reflects the selection pressure on it) will depend on whether or not character X has already been changed. This emphasizes an important point, namely that characters are to be considered independent only if they affect the fitness of the organism independently, the fitness effect of one not depending on the state of another. It is not sufficient that the genetic and physiological bases of the observed characters be independent.

Another way in which selection pressure can cause dependence is by a selection pressure affecting the same character in different species. For example, a series of cold years over a wide geographic area might make it probable that several phyletic lines would simultaneously become adapted to cold. If the probability of a cold wave in a particular time interval is p, the probabilities that zero, one, or two organisms in two lines become adapted to cold would be (1-p), 0, and p. If the occurrence of the cold wave is independent with respect to the two organisms, e.g., if they are widely separated geographically, the probabilities would be $(1-p)^2$, $2p(1-p)$, and p^2. An evolutionary tree requiring that both lines become adapted to cold at the same time would be more heavily discounted by the second model than by the first. This effect will occur only if the selection pressure varies in time. If it is constant, it will not cause dependence. Thus, if p is either 0 or 1, the two sets of probabilities given above are the same, being respectively 1, 0, 0 and 0, 0, 1. In the characters used by Cavalli-Sforza and Edwards (1967), they are able to show that selection variable in time but constant over space (and hence over phyletic lines existing at a given time) will not affect their estimation procedure. Unless we have fossils available, it may be impossible to detect directional changes affecting all of the organisms under study equally. We are more interested in the forces which tend to differentiate organisms than in those which cause parallel evolution when we set out to estimate phylogeny.

In all of the above cases, it is easier to point out the dangers of dependence of characters than to remedy them. The ideal solution is to consider the entire phenotype of the organism as a single character with a complex of states. The dependence of different characters in the same population is then reflected in the probabilities of change of this character. Although this is impractical in real cases, the combination of dependent characters with each other can often be carried out to a smaller extent. For example, if we have recorded the length and width of the skull of a mammal, and if shape stays approximately constant in evolution while size changes, both dimensions will change in a correlated way. But if we recode the data as size and length/width ratio, these characters will be independent. This type of recoding should be attempted whenever possible. Even when it is feasible, it will only correct for dependence within a species, and not for the effects of competition or of variation of the selection pressures.

Change in the Environment

I have mentioned above three reasons why a deterministic process of evolution could give a probability model of evolution which was not deterministic: statistical sampling error in the data, uncertainty about the initial state of the evolution process, and uncertainty about the model of evolution. An additional source is uncertainty about the state of the environment, which can also be thought of as a type of uncertainty about the model of evolution. If the response to a change in the environment is strictly deterministic, so that all of the randomness in the model reflects randomness of the environment, then the phenotypic differentiation of the group will reflect this randomness. This suggests use of the state of the environment as part of the phenotype of an organism. One aspect of the environment which is correlated with many other is geographical location. When we include as part of a species' phenotype its distribution, we are in effect coding environment into the phenotype. As long as we can formulate a model for the change of the environment, there would seem to be no particular objection to this sort of procedure.

An example of the duality between environmental randomness and random response to the environment is provided by Cavalli-Sforza and Edwards (1967). They consider a model of "selective drift" in which the selection coefficients in the equation for change of gene frequencies are assumed to be drawn at random from a statistical distribution. The randomness involved could be environmental randomness or could result from variability in the determination of a phenotype, which might then be selected deterministically.

Adaptive and Conservative Characters

The classical controversy over whether adaptive or non-adaptive characters should be used to establish phylogeny is rendered obsolete by a statistical approach. Any character for which a probability model can be constructed can be used, regardless of whether the basis of the probability is selection in a variable environment or genetic drift. Sokal and Sneath (1963) and Simpson (1961) point out that truly non-adaptive characters are rare if not non-existent, rendering the controversy somewhat moot if the definitions are interpreted rigidly. However, it seems likely that the intent of the taxonomists was to exclude characters whose similarity in different species reflected similarity of environment rather than phylogenetic history. A statistical inference approach de-emphasizes these characters. If a character is very labile, it is not improbable that it will show a great amount of convergent and parallel evolution. Evolutionary trees will not be made improbable by the fact that they require convergence or parallelism of these characters, so that they will be of little importance in establishing the form of the trees.

A similar statement applies to the common contention that "conservative" characters should be used in inferring relationship. The same thought underlies the rule, and a statistical inference approach automatically takes care of the problem for the same reason.

Weighting of Characters

Both of these concepts are related to the concept of "weighting" of characters, since both criteria effectively de-weight certain characters by rejection. The presence of weighting in classical phylogenetic methods distinguishes them from both phenetic methods (Davis and Heywood, 1963; Sokal and Sneath, 1963) and numerical phylogenetic methods, as outlined in Chapter I (see especially Cain and Harrison, 1960; Camin and Sokal, 1965). It would seem that weighting appears in a statistical inference approach. For example, if origin of state X_1 of character X from its primitive state X_o is much less probable than origin of state Y_1 of character Y from its primitive state Y_o, a tree requiring two origins of X_1 and one of Y_1 will be less likely than a tree requiring two origins of state Y_1 and one of state X_1. Thus more compromises in tree form will be made to avoid convergence in X than in Y, so that transition from X_o to X_1 is in effect more heavily weighted than transition from Y_o to Y_1. Notice that weight attaches not to whole characters but to specific evolutionary events within the characters. The events which are heavily weighted are those which are least probable *a priori*. Their occurrence more than once will make a tree much less probable. A change can be assigned a high weight by lowering its assumed probability.

Similarity to Present Phylogenetic Practice

It should not be thought that adoption of a statistical inference approach would invalidate present phylogenetic practice, although it would certainly cause some changes. As is apparent above, a statistical inference approach provides support for such practices as weighting characters. It is at least possible that taxonomists have been applying maximum likelihood or Bayesian methods in practice, with models of evolution derived from their experience. This could easily lead to estimation of much higher power than can presently be achieved numerically. This point is made by Rogers, Fleming, and Estabrook (1967) with regard to "parsimony," but I feel that it is more likely that present taxonomic practice resembles Bayesian inference, especially in view of the prevalence of weighting of characters.

It is not necessary that an explicitly statistical approach will yield better results than a classical one, given the crudity of the models of evolution used in statistical approaches at present. The advantage of a statistical approach lies in its ability to consider all data fairly, and in its well-defined nature which would tend to promote greater clarity of thought in discussions of phylogenetic methods.

DISCRETE CHARACTERS

The Basic Model

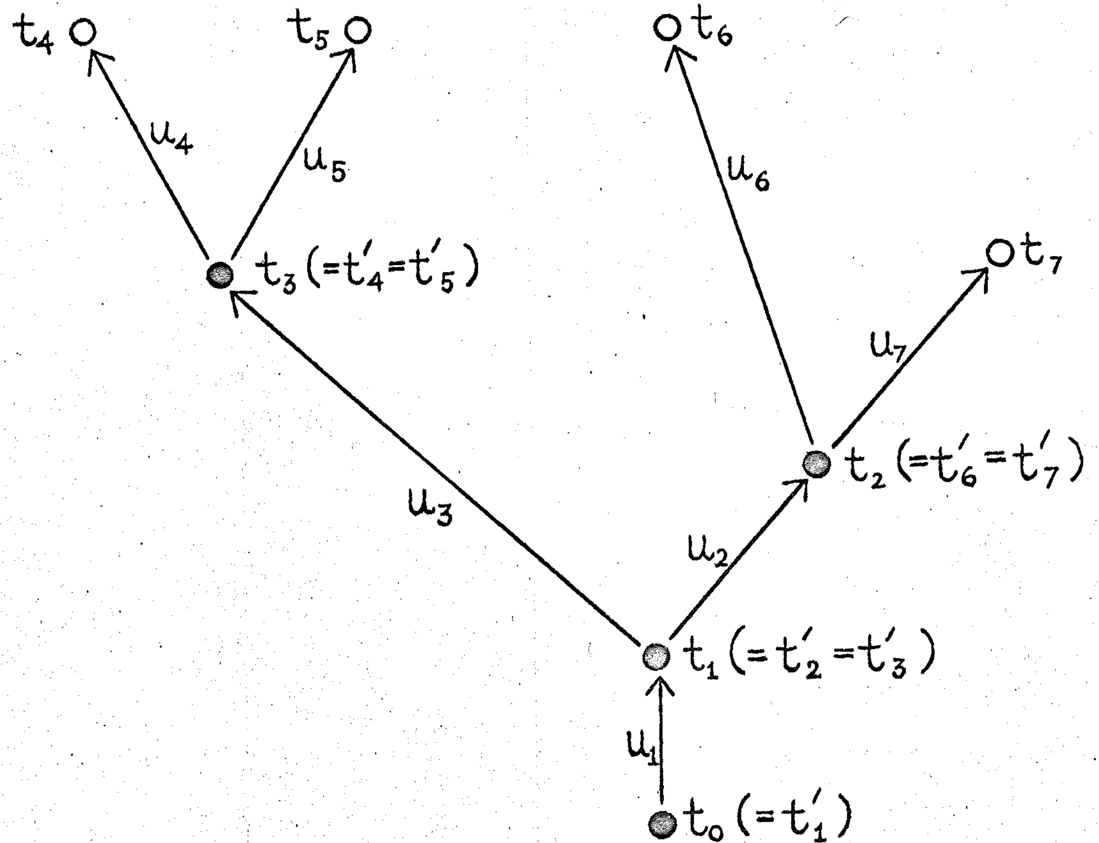
A discrete character is one whose phenotypes are distinct. For convenience we can number the phenotypes 1, 2, 3, . . . _k_, where _k_ is the number of phenotypes. In theory, the state of the character in a population could be characterized by giving the frequencies f_1, f_2, . . . , f_k of the _k_ phenotypes. But since we rarely know the genetic basis for the phenotypes, we cannot often make reasonable models for the change of these frequencies. It is also usually true that we have such small samples of each species that we cannot make decent estimates of their frequencies, although we may have an idea which phenotypes are present. In such cases, it seems reasonable to describe a population simply by listing the phenotypes it contains. If there are _k_ phenotypes, there are 2^k-1 such lists possible: (1), (2), (3), . . . , (k), (1,2), (1,3), . . . , (k-1,k), . . . , (1,2,3, . . . , k). Such a list describes the state of a character in a population. If polymorphism does not occur, we need only the first _k_ lists: (1), (2), . . . , (k). In the derivations which follow the possibility of polymorphism will be assumed.

Assume that the probability that a given change in the state of a character occurs in a given length of time depends only on the state of the character at the beginning of the interval and the length of the time interval. This means that the probability of any particular change cannot depend on the relative frequencies of the phenotypes comprising the state. Real evolutionary events consist of shifts in the frequencies of phenotypes, caused by selection, migration, and genetic drift, as well as the origin of new phenotypes by mutation and recombination. In our crude approximation to these processes, we have lost sight of all changes except those resulting in loss or addition of a phenotype. In mathematical terms, the above assumption states that each character undergoes an independent Markov process, in continuous time.

One additional assumption will be made which has no particular biological justification, but which simplifies the mathematics. Let the probability that there is no change in character *k* during a short time interval of length *dt* be $1 - w_k dt$, irrespective of the state of the character. All states of the character are then equally likely to be superseded by new ones. We now calculate the probability of a type II tree. Since the specifications of the tree include the data, this probability will be the *a priori* probability of the tree, rather than its likelihood. The length in time of segment *i* of the tree will be given by u_i, as in Fig. 4. The total length of all segments is $U = \Sigma u_i$. By the assumptions above, evolution in different characters is independent. Once the states of a character at the beginnings of two segments are given, evolution in the two segments is independent. If we can calculate the probability of the evolutionary events occurring in each segment of the tree, the probability of the tree will be the product of the segment probabilities. Consider character *j* in segment *i*. Let the states of the character at the beginning and end of the segment be *a* and *b*, these along with u_i being part of the specifications of the tree. Since w_j is the probability of change per unit of time, and since u_i is the length of the segment, the expected number of character state changes in the segment will be $w_j u_i$. The actual number of changes of state will be distributed according to a Poisson distribution with parameter $w_j u_i$. Then the *a priori* probability that there will be *k* changes of state is

Fig. 4.--Notation of times of branching on a tree of type II. Further explanation in text.

Figure 4



$$e^{-w_j u_i} (w_j u_i)^k / k! \qquad (1)$$

for any integral value of $\underline{k}$. Denote by $P_{ab}^{(k)}$ the probability that after $\underline{k}$ changes of state, the state is $\underline{b}$, given that the initial state was $\underline{a}$. Then the probability that after a length of time u_i the state is $\underline{b}$, given that it was initially $\underline{a}$ is

$$\sum_k P\,(\text{k changes occur in time } u_i)\; P_{ab}^{(k)}$$

which is

$$\sum_{k=0}^{\infty} e^{-w_j u_i} (w_j u_i)^k P_{ab}^{(k)} / k!$$

Multiplying this for all values of $\underline{i}$ and $\underline{j}$, we get for the probability of the tree:

$$P = \prod_i \prod_j \Big(\sum_k e^{-w_j u_i} (w_j u_i)^k P_{ab}^{(k)} / k! \Big). \qquad (2)$$

This expression is what I earlier referred to as $P(Q|T)$. It contains no term for the a priori probability of the form of the tree.

Given the specifications for the tree, expressed by the values of the u_i and the values of $\underline{a}$ and $\underline{b}$ (which are different for each character and each segment), and given the probability model of evolution, expressed by the w_j and the values of $P_{ab}^{(k)}$, we can calculate the probability of the tree by application of this formula. This formula is often laborious to use, especially when it must be calculated for many different trees in order to find the Bayesian estimate of the phylogeny. This is especially true when we are estimating trees of types III or IV. To each tree of type III, for example, there corresponds a large number of trees of type II, with different states of the character at the forks of the trees. The probability of a tree of type III is the sum of the probabilities of all of these trees of type II. This can be seen by considering the symbol Q, which represents the character states of all of the forks and branch tips of a tree of type II, to be composed of two parts: the states of the tips (D) and the states of the forks (F). We have an equation which allows us to calculate $P(F, D|T)$. For a tree of type III, we need to calculate $P(D|T)$, which is

$$\sum_{F} P(F, D|T)$$

so that we can calculate it by summing over all possible configurations F of the states of the fork points. For trees of type IV, consider the tree parameters represented by T to consist of parameters of the shape of the tree, $\underline{s}$, and parameters of the fork points, $\underline{t}$. We must calculate $P(D|s)$, which is

$$\sum_{F} \sum_{t} P(F, D, t|s),$$

and if the processes generating the form of the tree are independent of those changing the states of the characters:

$$P(D|s) = \sum_{t} \sum_{F} P(D, F|t, s)\ P(t|s)$$

so that if we know $P(t|s)$ or assume it to be equal for all values of the parameters $\underline{t}$, we can calculate $P(D|s)$ once we know $P(D, F|T)$ for all values of F. The estimation of trees of type I will not be discussed here, since complications arise, due to the fact that under almost any reasonable model of evolution any individual tree has probability zero, forcing us to deal with probability densities rather than probabilities.

"Parsimony" Methods and Bayesian Inference

Under certain assumptions, Bayesian methods are the same as "minimum evolution" or "parsimony" methods. First, let us find conditions under which the summation over k in equation (2) effectively consists of a single term. This will be the case if there is only one k for each i and j such that $P_{ab}^{(k)}$ is not zero, which is the same as saying that there is only one way to get from state a to state b, and that this way involves exactly k steps. If we assume that each character state can arise from only one other state, this condition is satisfied. The example in Fig. 1 has this property. If for a given character the value of w_jU (and hence of all w_ju_i) is very small, we can ignore all but one term in the summation in equation (2). For if there are two values of k for which $P_{ab}^{(k)}$ is nonzero, the term with the larger k will contain a higher power of w_ju_i, and will effectively vanish compared to the lower term.

Suppose that either one of these two conditions holds. Call the value of k in segment i for character j, n_{ij}. Equation (2) becomes

$$P = \prod_{ij} e^{-w_ju_i} (w_ju_i)^{n_{ij}} P_{ab}^{(n_{ij})} / n_{ij}! \ .$$

Let $W = \Sigma w_j$ and let $v_i = u_i/U$, so that $\Sigma v_i = 1$ and $u_i = v_iU$. Then

$$P = e^{-WU} \prod_{ij} (w_jU)^{n_{ij}} P_{ab}^{(n_{ij})} v_i^{n_{ij}}/n_{ij}! \ .$$

Taking natural logarithms

$$\log P = - WU + \sum_{ij} \log ((w_j U)^{n_{ij}} P_{ab}^{(n_{ij})})$$

$$+ \sum_{ij} n_{ij} \log v_i - \sum_{ij} \log (n_{ij}!). \qquad (3)$$

Let us assume that considering different evolutionary trees causes much larger change in the second term of (3) than in the other terms, so that log P is essentially determined by

$$\Sigma \log ((w_j U)^{n_{ij}} P_{ab}^{(n_{ij})}). \qquad (4)$$

This might result from a small value of the $w_j U$ or from large variation in $P_{ab}^{(n_{ij})}$ as a and b are altered. The summation in (4) can be considered as inversely related to the "amount of evolution." Its value depends on the n_{ij} and the a and the b, as well as on U, but not on the v_i. Thus it measures the type and number of evolutionary events, being insensitive to the finer details of their relationship to the lengths of the tree segments.

Suppose that the only allowable character state changes are those involving the gain or loss of one phenotype. Suppose further that the probability of a gain in $pw_j dt$ in a time interval of length dt, and that the probability of a loss is $qw_j dt$. Then if the change from a to b must involve at least n_G gains and n_L losses of phenotypes,

$$P_{ab}^{(n_{ij})} = p^{n_G} q^{n_L} m ,$$

where m is the number of different orderings of the gains and losses which are allowable, e.g., GGGLLL, GGLGLL, etc. Then

$$\sum_{ij} \log \left((w_jU)^{n_{ij}} P_{ab}^{(n_{ij})}\right)$$

$$= \sum_{ij} n_G \log (pw_jU) + \sum_{ij} n_L \log (qw_jU) + \sum_{ij} \log m \ .$$

The last term can be more or less ignored in comparison with the first two if w_jU is small. If an evolutionary tree requires n_G gains and n_L losses in segment _i_ for character _j_, we have

$$\log P \cong \sum_{ij} n_G \log (pw_jU) + \sum_{ij} n_L \log (qw_jU) \ . \qquad (5)$$

Thus each gain is assigned weight $-\log (pw_jU)$ and each loss is assigned weight $-\log (qw_jU)$, and the probability of a tree is inversely related to its weight. What all of the above boils down to is that if any change in a character is _a priori_ improbable over the time scale of the evolution of the group, we get a sufficient advantage from reducing the postulated number of changes that we can ignore the distribution of the changes among segments of the tree.

It may be objected that when working on recent groups, we usually have no idea of the time scale of evolution or of the rates of change of characters, so that we cannot say whether w_jU is small or large. If we attempt maximum likelihood estimation of U from equation (3), the estimate is $\hat{U} = N/W$, where $N = \Sigma n_{ij}$, so that $w_j\hat{U}$ will be Nw_j/W, where $W = \Sigma w_j$. If w_j is small compared to other w's, then $w_j\hat{U}$ will be small. In other words, if we have two sets of characters, with change being less probable in one set than in another, and if our data are such as to require about as many changes in one set as in the other, any tree in which there are less changes in the less labile characters at the expense of more changes in the more labile ones will have a higher probability, since the observed numbers of changes will then correspond more closely to their expectations.

Generally speaking, the above derivations give support to the use of "parsimony" methods only when we have advance knowledge of both the time scale of the evolution of the group (U) and the rates of change of the characters (as reflected by the values of w_j and $P_{ab}^{(k)}$), and these indicate that any change at all in the characters is unlikely. In this case, the less changes of character state must be assumed, the closer the number of changes becomes to its expectation.

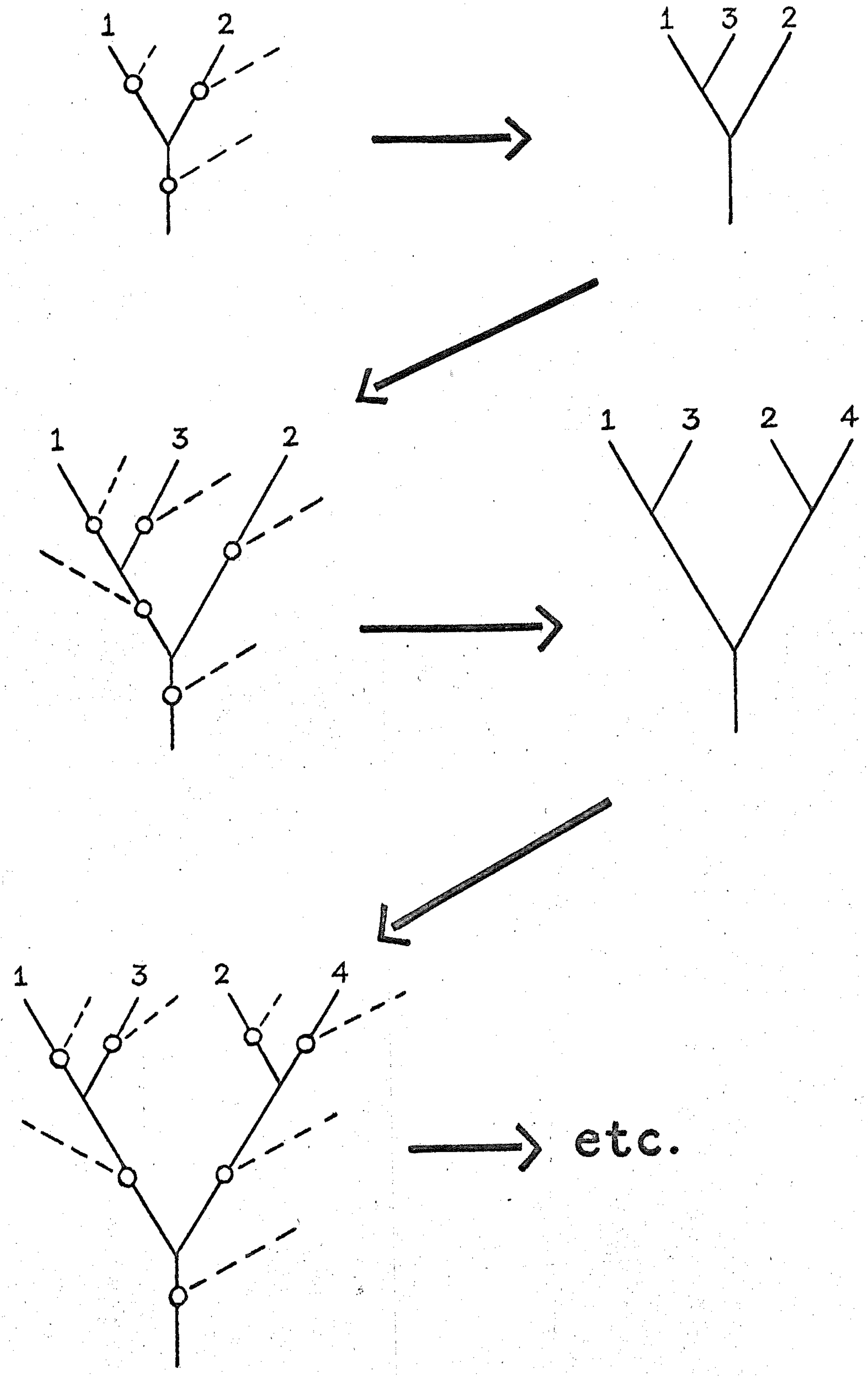
Use of "parsimony" approaches does, however, have the advantage of making estimation of type III and type IV trees simple. Each added character state change drastically lowers the probability. I have demonstrated that the probability of type III and type IV trees can be calculated by summing probabilities of type II trees. If one particular type II tree has the smallest number of character state changes, appropriately weighted, then only the type III and IV trees compatible with it have a reasonable probability, since the contribution that the "parsimonious" type II tree makes is so great compared to the contributions of the other type II trees to the other type III and IV probabilities.

It is relatively easy to develop algorithms to estimate type IV trees using parsimony models. The core of any such algorithm is a procedure which, given specifications for a tree of type IV, and given a set of data and information on the model of change in each character, calculates the probability of the tree, using equation (5). Obviously the optimum procedure would be simply to look at all possible trees and to pick the one with the highest value of log P (or, equivalently, the lowest total weight of character state changes). As we shall see below, the numbers of such trees are astronomical for even moderate numbers of species. In order to make the procedure practical we must examine only a small fraction of the total set of trees. A reasonable procedure is to make an initial guess at the best tree by some approximate procedure, then to try small alterations in the form of the tree, retaining those which improve the probability of the tree. When no further small alterations will improve the tree, we accept it as the estimate. This is by no means guaranteed to give the best tree. Several trials, using different initial guesses at the tree, are necessary to make the probability of obtaining the best tree moderately high. Of the results of these several trials, the one with the highest probability should be used as the final estimate.

The procedure of Camin and Sokal (1965) is of this type. They differ in the methods for making the initial guess. It is important that this guess be a good one, so that there is a high probability of arriving at the most "parsimonious" tree by the small alterations. The following procedure has been found to work fairly well. Let us confine interest to <u>bifurcating</u> trees, in which there are no forks giving rise to more than two branches. The procedure is illustrated in Fig. 5. We start by assigning a number to each species, in arbitrary order. Next combine species 1 and 2 into a two-species tree. Now we want to add species 3 to this tree. There are three places from which it could arise: below species 1, below species 2, or on the trunk of the tree. Try the species in each of these positions, and for each one compute the probability of the resulting three-species tree. Accept the position which has the highest probability. Adding the species has created two new segments of the tree. There are then five positions at which species 4 can be added. Try all five, calculating the probabilities of the resulting trees and choosing the position giving the highest probability.

Fig. 5.--The method of construction of the initial guess to a tree of type IV. Further explanation in text.

Figure 5



Continue in like manner, until all n species have been added. The resulting tree serves as the initial guess for the rearrangement routine. By renumbering the different species and carrying out the process, we get different initial guesses.

It can be shown that each bifurcating tree of type IV can be constructed by exactly one such process of stepwise addition of species. Since there are three positions to which the third species could be added, five to which the fourth species could be added, seven for the fifth species, and so on, the total number of trees of type IV is (1)(3)(5)(7) . . . (2n-3). This formula was derived by Cavalli-Sforza and Edwards (1967; also Edwards and Cavalli-Sforza, 1964). For n = 10 the number is 34,459,425, and for n = 20 it is more than 8×10^{21}. However, the above procedure for constructing the initial guess examines only 3 + 5 + 7 + 9 + . . . + (2n-3) trees. For n = 10 and n = 20 this is 80 and 360, respectively. Only a tiny fraction of all possible trees is examined.

CONTINUOUS CHARACTERS

Transformation of the Characters to Independence

When the characters recorded for a group of species are measurements on continuous scales, a different methodology applies. Assume that we have $\underline{p}$ variables observed in $\underline{n}$ species, each species being represented by several specimens. Suppose that the value of character $\underline{i}$ in a given individual is y_i. Assume that each character is determined additively by $\underline{p}$ genes. If z_j takes on the value 0, 1/2, or 1 according to whether locus $\underline{j}$ has genotype $a^{(j)}a^{(j)}$, $A^{(j)}a^{(j)}$, or $A^{(j)}A^{(j)}$, we can write the $\underline{i}$-th phenotype of the individual as:

$$y_i = \sum_j a_{ij} z_j + m_i \tag{1}$$

where m_i is the phenotype when all loci are homozygous for the $\underline{a}$ alleles. The expectation of z_j is, of course, the frequency q_j of the $A^{(j)}$ allele in the population in question. The variance of z_j between individuals is expected to be $2q_j(1-q_j)$. If the loci have random association of alleles (i.e., no "linkage disequilibrium") we have Cov $(z_k, z_\ell) = 0$ for all $\underline{k}$ and $\underline{\ell}$ in which $k \neq \ell$. Over a large range of gene frequencies q_j, $2q_j(1-q_j)$ will be near 1/2. For example, when $q_j = 1/2$ it is 1/2, while for $q_j = 0.3$ it is 0.42. Thus, approximately, Var $(z_i) = 1/2$. Then the variance-covariance matrix of the z_i is $V(Z) = (1/2)\,I$, where I is the identity matrix.

We can write (1) in matrix notation as $Y = A\,Z + M$, where Y is the vector of y_i's, A is the matrix of a_{ij}'s, Z is the vector of z_i's, and M is the vector of m_i. It is easily shown that the variance-covariance matrix of the Y is

$$V(Y) = A\ V(Z)\ A^t = (1/2)\ A\ A^t\ .$$

(All capital letters are matrices or vectors unless otherwise noted.) We can observe the variance-covariance matrix of Y, but not the matrix A. V(Y) will be positive definite, since it is a variance-covariance matrix. Every such matrix can be written in the form $V(Y) = U\ L\ U^t$, where U is orthogonal and L is diagonal. Since the components of L are all positive, we can take their square roots, obtaining the matrix $L^{1/2}$, and the reciprocals of their square roots to get $L^{-1/2}$. Let us take $B = L^{1/2}\ U^t$. If we apply the linear transformation B to Y, we get $X = B\ Y$, where X has the variance-covariance matrix

$$\begin{aligned} V(X) &= B\ V(Y)\ B^t \\ &= (L^{-1/2}\ U^t)\ U\ L\ U^t\ (U\ L^{-1/2}) \\ &= I \end{aligned}$$

since $U^tU = I$. Thus, knowing V(Y), which can be estimated from the within-species covariances, we can apply a linear transformation B to obtain a vector X whose entries are independent within populations. This is a standard procedure in multivariate statistical analysis, forming part of the process of canonical analysis (see Seal, 1964). Thus any set of data can be transformed into one in which there is no within-species covariance of characters. The new variables X are still linear combinations of the underlying variables Z, i.e., $X = (BA)\ Z$. The transformation BA has the property that $(BA)(BA)^t = 2I$.

The gene frequencies q_i can be treated approximately as if they were undergoing a random walk or Brownian motion with variance per generation

$q_i(1-q_i)/2N \cong 1/8N$, where N is the effective population size, assumed the same for all species. Then if Q(t) is the vector of gene frequencies at time t, approximately

$$E\ (Q(t+s) - Q(s)) = 0$$
$$V\ (Q(t+s) - Q(s)) = (t/8N)\ I.$$

The distribution of Q being approximately multivariate normal:

$$Q(t + s) - Q(s) \sim N_p\ (0,\ (t/8N)I).$$

If time is measured in units of 4N generations,

$$Q(t + s) - Q(s) \sim N_p(0,\ (t/2)\ I).$$

Consider the mean phenotype vectors $\overline{X}(t+s)$ and $\overline{X}(s)$. We have Q = E(Z) so $\overline{X}$ = (BA) Q and

$$\overline{X}(T + s) - \overline{X}(s) = (BA)\ (Q(t+s) - Q(s))$$

so $\overline{X}(t+s) - \overline{X}(s)$ is normally distributed with mean 0 and variance

$$(BA)\ (t/2)\ I(BA)^t,$$

so that $\overline{X}(t+s) - \overline{X}(s) \sim N_p(0,\ tI)$. Henceforth, I leave the bar off the X(t). Thus the mean phenotypes $\overline{x}_i$ can be treated as if they were performing independent Brownian motion processes with variance 1/4N per generation (or 1 per 4N-generation unit).

Estimation of Trees

Turning to the question of estimation of type II trees, we are in the same situation as Cavalli-Sforza dnd Edwards (1967). There is no need to repeat their definitive analysis here, except to note that it is applicable. If one is estimnting trees of type II, one has specified the vectors $X^{(k)}(t_i)$ at each of the forks of the tree, so that the probability density of the whole tree is

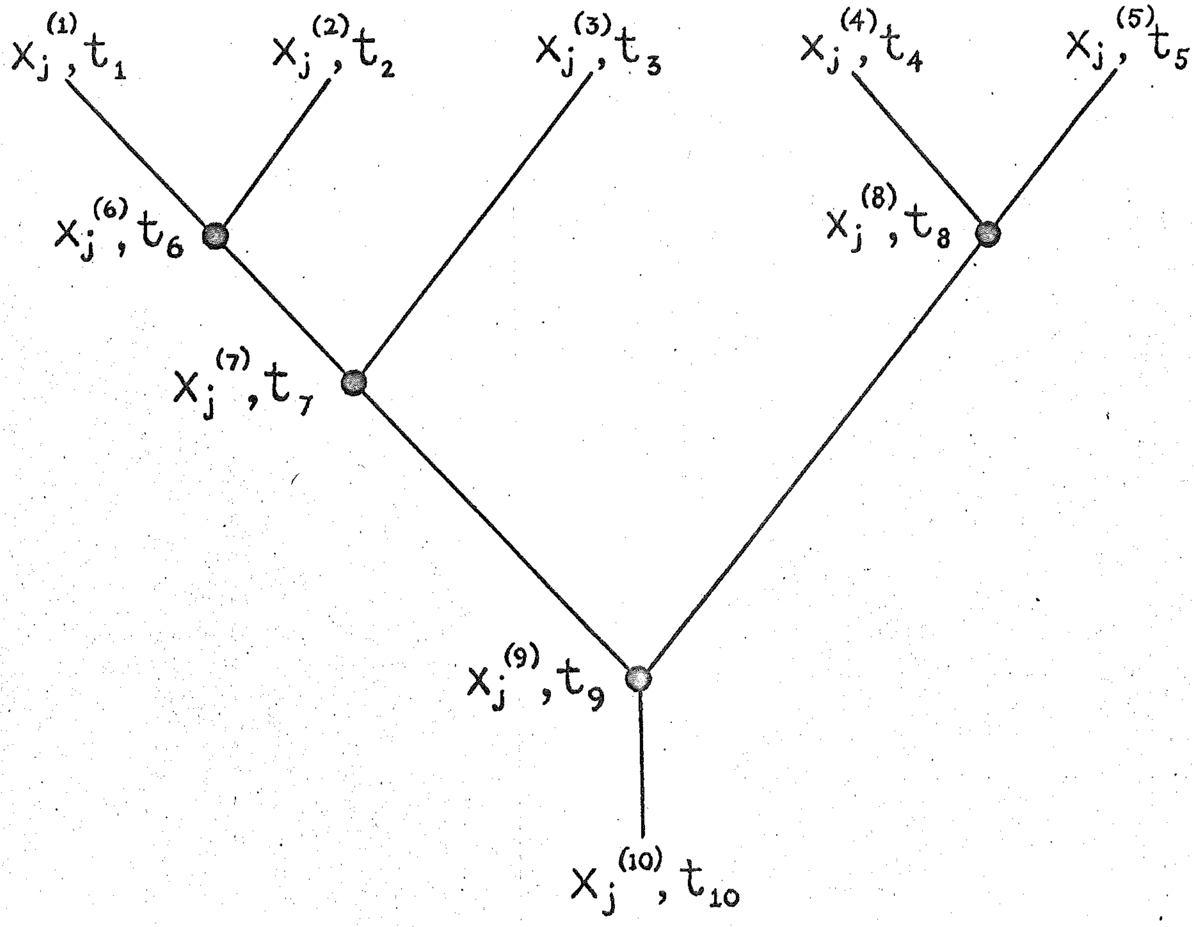
$$P = \prod_i \frac{1}{(2\pi)^{np/2}(t_i - t_k)^{np/2}} \exp\,(-\sum_j (x_j^{(i)}(t_i) - x_j^{(k)}(t_k))^2$$

$$/(t_i - t_k))\, dx_1^{(1)} \; . \; . \; . \; dx_j^{(i)} \; . \; . \; . \; dx_n^{(p)}$$

where <u>k</u> is the fork ancestral to <u>i</u>. This is readily computed for any tree. Cavalli-Sforza and Edwards point out some difficulties which arise in estimation of the times of branching, t_i. They maintain that, because of the existence of singularities in the likelihood surface, methods other than maximum likelihood must be used to guess that t_i. It should be mentioned at this point that although Cavalli-Sforza and Edwards refer to their procedure as maximum-likelihood estimation, it is actually Bayesian inference. As pointed out in the General Principles section, the specifications for a tree of type II contain the data, hence the likelihood is either 0 or 1, and we must maximize the prior probability over all trees with likelihood 1.

It should be evident that when Cavalli-Sforza and Edwards use estimation procedures to obtain a tree of type II and then present the results as a tree of type III (Cavalli-Sforza and Edwards, 1965) or type IV (Edwards and Cavalli-Sforza, 1964), there is some loss of power in the estimation. If they did not attempt to estimate fork phenotypes $X(t_i)$ they would presumably get a better type III tree. It is obviously of interest to find ways to compute the probabilities of type III trees. Suppose that we have the tree pictured in Fig. 6. Let <u>i'</u> be the number of the fork below point <u>i</u> (so that if <u>i</u> = 1, Fig. 6.--Notation of times of branching and mean phenotypes on a tree of type II with continuous characters. Further explanation in text.

Figure 6


$x_j^{(1)}, t_1$
$x_j^{(2)}, t_2$
$x_j^{(3)}, t_3$
$x_j^{(4)}, t_4$
$x_j^{(5)}, t_5$
$x_j^{(6)}, t_6$
$x_j^{(8)}, t_8$
$x_j^{(7)}, t_7$
$x_j^{(9)}, t_9$
$x_j^{(10)}, t_{10}$

$\underline{i}'=6$, etc.). Now notice that for character $\underline{j}$

$$x_j^{(1)} = (x_j^{(1)} - x_j^{(6)}) + (x_j^{(6)} - x_j^{(7)}) + (x_j^{(7)} - x_j^{(9)}) + (x_j^{(9)} - x_j^{(10)}) + x_j^{(10)}$$

$$x_j^{(2)} = (x_j^{(2)} - x_j^{(6)}) + (x_j^{(6)} - x_j^{(7)}) + (x_j^{(7)} - x_j^{(9)}) + (x_j^{(9)} - x_j^{(10)}) + x_j^{(10)}$$

$$x_j^{(3)} = (x_j^{(3)} - x_j^{(7)}) + (x_j^{(7)} - x_j^{(9)}) + (x_j^{(9)} - x_j^{(10)}) + x_j^{(10)}$$

etc.

By the model of evolution $x_j^{(i)} - x_j^{(i')}$ and $x_j^{(k)} - x_j^{(k')}$ are independent (if $i \neq k$), since they measure Brownian motion of different particles or in different time intervals. Let $a_i = x^{(i)} - x^{(i')}$ for a given character. Then $E(a_i) = 0$, and if $u_i = t_i - t_{i'}$,

$$\text{Var }(a_i) = t_i - t_{i'} \quad \text{and Cov }(a_i, a_j) = 0 \text{ if } i \neq j.$$

The variables a_i are normally distributed. Then the $x_j^{(i)}$, being sums of normal variables, are normally distributed. However, they are not independent. For example,

$$\text{Cov }(x_j^{(1)}, x_j^{(3)}) = \text{Cov }(a_1 + a_6 + a_7 + a_9 + x_j^{(10)}, a_3 + a_7 + a_9 + x_j^{(10)})$$

since Cov $(a_i, a_j) = 0$ and $x_j^{(10)}$ is a constant:

$$\mathrm{Cov}\ (x_j^{(1)},\ x_j^{(3)}) = \mathrm{Var}\ (a_7) + \mathrm{Var}\ (a_9)$$

$$= t_7 - t_9 + t_9 - t_{10} = t_7 - t_{10}\ .$$

In general, if the lines leading to i and j originate at t_o and split from each other at t_{ij},

$$\mathrm{Cov}\ (x_k^{(i)} - x_k^{(j)}) = t_{ij}\ .$$

Given a type III tree, we can calculate t_{ij} for every i and j. Let T be a matrix whose elements are the t_{ij}. Then the vector X has $E(X) = 0$ and $V(X) = T$, so that the probability density of the tree is

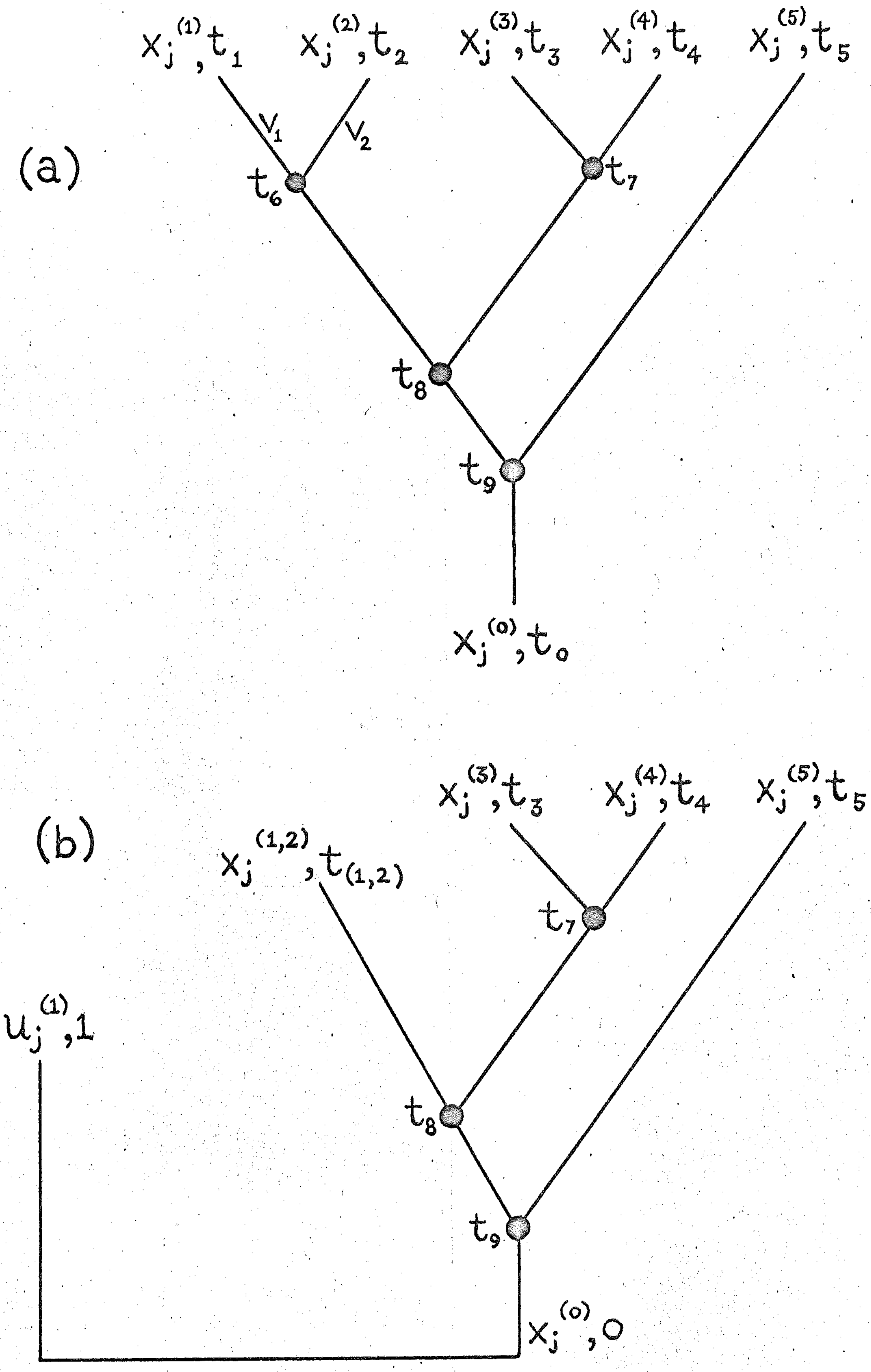
$$P = \frac{1}{(2\pi)^{p/2}\ \det\ (T)}\ \exp\ (\ -\ X^t\ T^{-1}\ X)\ .$$

Given the vector X and the values of T, this can be computed for any tree. However, it requires a matrix inversion and the evaluation of a determinant.

Computation of P More Rapidly

A way of computing P which enables a great amount of time to be saved is to transform the shape of the tree, all transformations holding P constant, until T becomes a diagonal matrix. The variables $x_j^{(i)}$ will be transformed to the variables $u_j^{(i)}$ such that the $u_j^{(i)}$ are independent and each has unit variance. Consider the tree in Fig. 7(a). On every such tree we can find at least two branch tips which are adjacent. In this case we can choose tips 1 and 2. Consider the transformation which replaces $x_j^{(1)}$ and $x_j^{(2)}$ by $x_j^{(1)} - x_j^{(2)}$ and a variable of the form $c\ x_j^{(1)} + (1 - c)\ x_j^{(2)}$. The value of c which will be used will be determined later. We have for $k \neq 1, \neq 2$, Fig. 7.--Illustration of the method for rapid calculation of the probability of a type III tree with continuous characters. Further explanation in text.

Figure 7


(a)
$x_j^{(1)}, t_1$
$x_j^{(2)}, t_2$
$x_j^{(3)}, t_3$
$x_j^{(4)}, t_4$
$x_j^{(5)}, t_5$
v_1
v_2
t_6
t_7
t_8
t_9
$x_j^{(0)}, t_0$
(b)
$x_j^{(3)}, t_3$
$x_j^{(4)}, t_4$
$x_j^{(5)}, t_5$
$x_j^{(1,2)}, t_{(1,2)}$
t_7
$u_j^{(1)}, 1$
t_8
t_9
$x_j^{(0)}, 0$

$$\mathrm{Cov}\ (x_j^{(k)},\ x_j^{(1)} - x_j^{(2)}) = \mathrm{Cov}\ (x_j^{(k)},\ x_j^{(1)})$$

$$- \mathrm{Cov}\ (x_j^{(k)},\ x_j^{(2)}) = 0$$

so that $x_j^{(1)} - x_j^{(2)}$ is independent of $x_j^{(3)}, \ldots, x_j^{(n)}$. It has variance

$$\mathrm{Var}\ (x_j^{(1)} - x_j^{(2)}) = \mathrm{Var}\ (x_j^{(1)}) + \mathrm{Var}\ (x_j^{(2)})$$

$$- 2\ \mathrm{Cov}\ (x_j^{(1)},\ x_j^{(2)}) = t_1 + t_2 - t_6\ .$$

Our new variate will be

$$u_j^{(1)} = (x_j^{(1)} - x_j^{(2)})/(t_1 - t_6 + t_2 - t_6)^{1/2}$$

which will have expectation 0 and variance 1. If it is to be independent of $c\ x_j^{(1)} + (1 - c)\ x_j^{(2)}$ we must have

$$\mathrm{Cov}\ (x_j^{(1)} - x_j^{(2)},\ c\ x_j^{(1)} + (1 - c)\ x_j^{(2)}) = 0$$

so that

$$c\ t_1 - c\ t_6 + (1 - c)\ t_6 - (1 - c)\ t_2 = 0$$

whereby $\quad c = (t_2 - t_6)/(t_1 - t_6 + t_2 - t_6).$

Write v_1 for $t_1 - t_6$ and v_2 for $t_2 - t_6$. The variance of $c\ x_j^{(1)} + (1-c)\ x_j^{(2)}$ will be

$$\mathrm{Var}\ (c\ x_j^{(1)} + (1 - c)\ x_j^{(2)}) = c^2\ t_1 + (1 - c)^2\ t_2 + 2c(1 - c)\ t_6$$

$$= (v_2^2 t_1 + v_1^2 t_2 + 2 v_1 v_2 t_6)/(v_1 + v_2)^2 = t_6 + v_1 v_2/(v_1 + v_2)$$

and its covariance with every other $x_j^{(k)}$ will be

$$\begin{aligned}
&\mathrm{Cov}\ (c\ x_j^{(1)} + (1 - c)\ x_j^{(2)},\ x_j^{(k)}) \\
&\qquad = c\ \mathrm{Cov}\ (x_j^{(1)},\ x_j^{(k)}) + (1 - c)\ \mathrm{Cov}\ (x_j^{(2)},\ x_j^{(k)}) \\
&\qquad = \mathrm{Cov}\ (x_j^{(1)},\ x_j^{(k)}) = \mathrm{Cov}\ (x_j^{(2)},\ x_j^{(k)}).
\end{aligned}$$

Thus if we make the transformation

$$u_j^{(1)} = (x_j^{(1)} - x_j^{(2)})/(v_1 + v_2)^{1/2}$$

and

$$x_j^{(1,2)} = (v_2\ x_j^{(1)} + v_1\ x_j^{(2)})/(v_1 + v_2)$$

$u_j^{(1)}$ will be independent of the other variables, with expectation 0 and variance 1, while $x_j^{(1,2)}$ will be distributed as if it were the phenotype of a branch tip which is situated beyond point 6, at time

$$t_6 + v_1\ v_2/(v_1 + v_2).$$

Then we have altered the tree in Fig. 7(a) to that in Fig. 7(b). Once this is done, we can choose two adjacent tips in the new tree and repeat the process (in this case with $x_j^{(3)}$ and $x_j^{(4)}$). The procedure is continued until p - 1 variables $u_j^{(1)}$, $u_j^{(2)}$, . . . , $u_j^{(n-1)}$ have been computed. There will remain a residual variable (which we can call x_j') which will be expected to have expectation $x^{(0)}$ and a variance given by the value of t assigned to it by the above process, say t'. Then we compute $u_j^{(p)} = (x_j' - x_j^{(0)})/t'$. The probability density of the transformed variables is now easily computed to be

$$P = \frac{1}{(2\pi)^{np/2}}\ \exp\ (-\ 1/2 \sum_{ij}\ (u_j^{(i)})^2)\ du_1^{(1)}\ .\ .\ .\ du_{n-1}^{(p)}.$$

If we do not know the value of $x_j^{(0)}$, its maximum likelihood estimate is obviously x_j', so that $u_j^{(p)}$ will be 0.

Thus, given a set of values of the $x_j^{(i)}$ and knowing the t_i, we can compute a set of $u_j^{(i)}$ and evaluate P. But since the t_i enter into the computation of the $u_j^{(i)}$ in fairly complicated ways, it is not easy to see how to compute the Bayesian estimates of the t_i, given a shape of tree and the $x_j^{(i)}$. Note that the transformation from x's to u's must be done once for each character (for each value of j). There is an easier way of taking all of the characters into account. Note that the characters x_j are transformed from the original set of data y_j. It can be shown that Mahalanobis' Distance between species i and k is

$$D_{ik}^2 = \sum_{j\ell} y_j^{(i)} ((V(Y))^{-1})_{j\ell}\, y_\ell^{(k)} = \sum_m (x_m^{(i)} - x_m^{(k)})^2 .$$

I will not give the proof here, but it can be shown that if we follow the algorithm below, P will be calculated properly:

(a) Calculate all D_{ik}^2 from the original data.

(b) Remove two adjacent tips, calculating $U_1^2 = D_{12}^2/(v_1 + v_2)$.

(c) Create a new "species" (1,2), calculating its Mahalanobis distances with the remaining species as

$$D_{(1,2),j}^2 = D_{1j}^2\, v_2/(v_1 + v_2) + D_{2j}^2\, v_1/(v_1 + v_2)$$
$$- D_{12}^2\, v_1 v_2/(v_1 + v_2)^2$$

and its time of occurrence as $t_{(1,2)} = t_6 + v_1 v_2/(v_1 + v_2)$.

(d) Continue until p - 1 values, $U_1^2, U_2^2, \ldots, U_{p-1}^2$ have been calculated. Then calculate the probability of the tree as

$$P = (2\pi)^{-np/2} \exp\left(- \sum_i U_i^2/2\right).$$

These procedures enable calculation of P for any hypothesized tree. It seems likely that they can be used to estimate trees of type III without any great difficulties. It is not clear without further examination whether the singularities which plagued Cavalli-Sforza and Edwards (1967) will appear here.

Models of Evolution involving Natural Selection

So far the model of evolution has been assumed to be one of random genetic drift. It is also possible to justify the same estimation procedures by models involving selection. Cavalli-Sforza and Edwards (1967) have presented a model of "selective drift" in which the Brownian motion of the x_j is due to fluctuations of selective value. Their x_j are, of course, allele frequencies. In the above case the x_j are phenotypes rather than allele frequencies. However, a similar model can be proposed. Let $x_j(t + 1) - x_j(t)$ represent the change in x_j in one generation. Then, approximately,

$$\overline{x}_j(t + 1) - \overline{x}_j(t) = s(t)\ h^2\ (\mathrm{Var}\ (x_j(t))^{1/2}$$

where h^2 is the fraction of within-population variance of x_j which is additive genetic variance, and $s(t)$ is a selection coefficient which is the change in mean fitness per standard deviation of change in x_j. If $s(t)$ is a random variable drawn from a normal distribution, and $E(s(t)) = 0$, with no autocorrelation between successive values of s, then $x_j(t + w) - x_j(w)$ is also normally distributed with variance proportional to t and to the variance of s. The variables x_j will perform independent Brownian motions in time and the process will have the same properties as the random genetic drift model, except that time must be scaled in units of $2/(h^2 \text{ Var } (s))$ generations instead of units of 4N generations, with h^2 and Var (s) assumed to be the same for all characters. A cruder model can also be set up in which it is assumed that the changes in x_j are due to gene substitutions, that the effect of a substitution is equally likely to increase or decrease x_j, and that the size of this change is proportional to the standard deviation of x_j within a population (which might be the case if all of the x_j were controlled by the same number of loci). The behavior of this model would be the same as that of the other two, except that the time scale is again different. The difficulty with the last two models is that selection must act independently on the x_j rather than on the observed characters y_j, or on some underlying variables correlated with the x_i. It might be possible to construct more realistic models under which the x_j would perform independent Brownian motion. The situation certainly bears examination.

Pretransformation of the Variables

A basic assumption throughout this section has been that the genes controlling the variation of the characters affect them additively. This leads to some difficulty when the characters are measurements of lengths, areas, or volumes in an individual. All of these have a lower limit of 0. In addition, all are affected by the size of the organism. If S is the size of an organism, linear measurements are expected to be proportional to S, areas to S^2, and volumes to S^3. This will lead to correlation of the characters. Furthermore, the correlation cannot be removed by linear transformations such as the one which produces the x_j from the y_i. However, if we take the logarithms of the original measurements and use them as the characters, the situation is greatly simplified. The characters can now range from

$-\infty$ to $+\infty$, as is proper for normally-distributed variables, and size enters additively instead of multiplicatively. If a measurement w has an allometric equation $w = a S^b$, where S is size, then $\log w = \log a - b \log S$. Under these circumstances the transformed variables x_i will include one variable which measures size, and the others, being independent of it, will be "shape" variables. Thus it is desirable to take logarithms of measurements before including them in this type of analysis.

SUMMARY

Classical methods for the construction of phylogenies are reviewed briefly and found to be either ill-defined or ill-justified. It is suggested that the inference of phylogenies be viewed as a problem in statistical inference, with the model of evolution which is assumed supplying the probabilities of the data given the phylogenies. Four types of phylogenetic trees are defined, and some of the general difficulties encountered in estimating different types of trees by the methods of Bayesian inference and maximum likelihood are discussed. The use of data from fossil organisms and the weighting of characters in phylogenetic inference are discussed from this point of view.

A model of evolution of characters which have discrete states is stated, and a general expression for the likelihood of a phylogenetic tree is derived. The formula is then used to examine when Bayesian methods will give results identical with methods of "minimum evolution." The models and results of Cavalli-Sforza and Edwards are extended, and methods for rapid calculation of the likelihood of an evolutionary tree are developed.

REFERENCES

Blackwelder, R. E. 1967. Taxonomy. John Wiley, New York.

Cain, A. J., and G. A. Harrison. 1960. Phyletic weighting. Proc. Zool. Soc. London 135:1-31.

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. 1965. Analysis of human evolution, pp. 932-951. In S. J. Geerts [ed.], Genetics Today. Pergamon Press, Oxford.

________. 1967. Phylogenetic analysis--models and estimation procedures. Amer. J. Human Genet. 19:233-257.

Davis, P. H., and V. H. Heywood. 1963. Principles of Angiosperm Taxonomy. D. Van Nostrand, Princeton.

Doolittle, R. F., and B. Blombäck. 1964. Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. Nature 202:147-152.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees, pp. 67-76. In V. H. Heywood and J. McNeill [eds.], Phenetic and Phylogenetic Classification. Systematics Association Publ. No. 6, London.

Farris, James S. 1967. The meaning of relationship and taxonomic procedure. Systematic Zoology 16:44-51.

Feller, W. 1957. An Introduction to Probability Theory and Its Applications. Vol. I. 2d ed. John Wiley, New York.

Hennig, Willi. 1967. Phylogenetic Systematics. Univ. of Illinois Press, Urbana.

Mayr, E., E. G. Linsley, and R. L. Usinger. 1953. Methods and Principles of Systematic Zoology. McGraw-Hill, New York.

Rogers, D. J., H. S. Fleming, and George Estabrook. 1967. Use of computers in studies of taxonomy and evolution. Evolutionary Biol. 1:169-196.

Seal, H. 1964. Multivariate Statistical Analysis for Biologists. John Wiley, New York.

Sharrock, G. 1968 (Manuscript in preparation)

Sokal, R. R., and P. H. A. Sneath. 1963. Principles of Numerical Taxonomy. W. H. Freeman, San Francisco.

Throckmorton L. H. 1962. The problem of phylogeny in the genus Drosophila. Studies in Genetics (Univ. of Texas Publ. 6205) 2:207-343.

________. 1965. Similarity versus relationship in Drosophila. Systematic Zool. 14:221-236.

Wagner, W. H. 1961. Problems in the classification of ferns, pp. 841-844. In Recent Advances in Botany. Vol. I. Univ. of Toronto Press, Toronto.