TOWARD DEFINING THE COURSE OF EVOLUTION: MINIMUM CHANGE FOR A SPECIFIC TREE TOPOLOGY

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. of Physiological Chemistry, Univ. of Wisconsin, Madison, Wisconsin, 53706), 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool., 20:406–416.—A method is presented that is asserted to provide all hypothetical ancestral character states that are consistent with describing the descent of the present-day character states in a minimum number of changes of state using a predetermined phylogenetic relationship among the taxa represented. The character states used as examples are the four messenger RNA nucleotides encoding the amino acid sequences of proteins, but the method is general. [Evolution; parsimonious trees.]

It has been a goal of those attempting to deduce phylogenetic relationships from information on biological characteristics to find the ancestral relationship(s) that would permit one to account for the descent of those characteristics in a manner requiring a minimum number of evolutionary steps or changes. The result could be called the most parsimonious evolutionary tree and might be expected to have a high degree of correspondence to the true phylogeny (Camin and Sokal, 1965). It's justification lies in the most efficient use of the information available and does not presuppose that evolution follows a most parsimonious course. There are no known algorithms for finding the most parsimonious tree(s) apart from the brute force method of examining nearly every possible tree.[1] This is impractical for trees involving a dozen or more taxonomic units. Most numerical taxonomic procedures (Sokal and Sneath, 1963; Farris, 1969, 1970; Fitch and Margoliash, 1967) provide dendrograms that would be among the more parsimonious solutions; one just cannot be sure that a more parsimonious tree structure does not exist. Farris (1970) has explicitly considered the parsimony principle as a part of his method which, like the present method, has its roots in the Wagner tree (Wagner, 1961, 1965).

A problem subordinant to finding the most parsimonious tree(s) is finding all possible, most parsimonious assignments of the information to any given particular tree. This is the problem that is treated herein. Its solution provides the most reasonable hypotheses on the ancestral states and therefore is the best estimate of the course of evolution. The biological characteristics to be used are the nucleotides of orthologous[1] genes but the method is applicable to any data for which the underlying assumptions are acceptable. Thus, given a set of descendent nucleotide sequences and a topology presumed to describe their ancestral relationships one can set forth: a), the exact number of nucleotide replacements that are the minimum necessary to account for the descent of those sequences from a common ancestor; b), all possible ancestral

[1] An elegant beginning to an attack on the problem has recently been published by Farris (1969) who developed a method which estimates the reliability of various characters and then weights the characters on the basis of that reliability.

[1] Orthologous is used, as previously defined (Fitch, 1970), to denote that particular subset of homologous genes for which there is an exact one-to-one correspondence (ortho = exact) between the ancestral relationships of the genes and the ancestral relationship of the species (or other taxonomic unit) in which those genes are found. It may be distinguished from paralogous which is used to denote homologous genes that arose via a gene duplication and descended side by side (para = in parallel) in any one line of descent. An example of the latter would be the genes for α and β hemoglobin.

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

sequences at each node (branch point) that are consistent with that minimum; c), those combinations of nodal ancestral sequences that are consistent with that minimum number of replacements; and d), a way of assigning relative weights among the various alternative nucleotide replacements required of the various internodal periods of the several minimal trees.

METHOD

The method assumes that: a), a set of orthologous descendent sequences is available; b), the ancestral (topological) relationships among them are known; and c), any nucleotide may replace any other nucleotide at any position without regard to the nature of prior replacements in that position or elsewhere. The word "may" indicates that all logical options are to be considered; this is not to deny the possibility that selective forces may have precluded some of those options.

The character state of a given nucleotide position is the set of nucleotides that might be present, A, C, G or U or some combination thereof. Ambiguity may arise because one is uncertain which codon is used in the gene. For example, arginine may be encoded by either AGG or CGG (among others) and we can not be sure, knowing only that arginine was encoded, whether the first nucleotide position is A or C. Consequently, the character state is represented by the set AC to indicate the exact range of ambiguity. Ambiguity will also arise from uncertainty about ancestral character states and will be similarly represented.[1]

The presentation will examine only a single nucleotide position in all the taxonomic units. In practice, the procedure given here must be repeated for every nucleotide position in the gene. The reconstruction of the ancestral nucleotide(s) then follows two phases, the preliminary phase in which nucleotides are placed on all ancestral nodes, and a final phase, in which corrections to the preliminary assignments are made. The procedure to this point has many resemblances to that of Farris (1970), their differences being attributable to differences in the nature of the phenetic data and in the assumptions regarding the way in which the character states may change. It will sometimes occur where the nucleotide assignments are ambiguous that certain of the nucleotides assigned to successive ancestral nodes can not both be part of the same tree of minimum evolution, i.e., they can not validly be linked to form a part of a most parsimonious tree. Therefore, following the sections on reconstructing the ancestral nucleotide sets, is a section that shows which links between nucleotides at successive nodes are valid components of a complete linkage in a tree requiring the minimum number of nucleotide replacements to account for its evolution. Finally, since it would appear that not every possible link between two successive nodes should be treated as equally representative of those events that did occur in the descent, there follows a section on determining the relative "probabilities" that the different permitted links may have occurred in the descent of these nucleotides. The quotation marks on probabilities will be explained at that time.

[1] Another type of ambiguity concerns missing information which may be missing for one of two reasons, viz. either the character was not examined or the character does not exist. These suggest different treatments. Where the character was not examined, it may be best to represent the state as completely ambiguous (e.g., ACGU) or, in set theoretic terminology, as the universal set. The result is that the character has no influence whatsoever on the procedures described or on the results obtained. Where the character does not exist, it may be best to represent the state by the empty (null) set. The result is that a change of state will be detected (viz. the deletion of the character) but there is no other effect on the procedures or results. One must exercise some discretion, however, since, if five contiguous codons were eliminated in a single genetic event by unequal crossingover and all 15 (3×5) nucleotide positions assigned to empty set state, the final result would show 15 changes of state where only one deletion occurred.

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

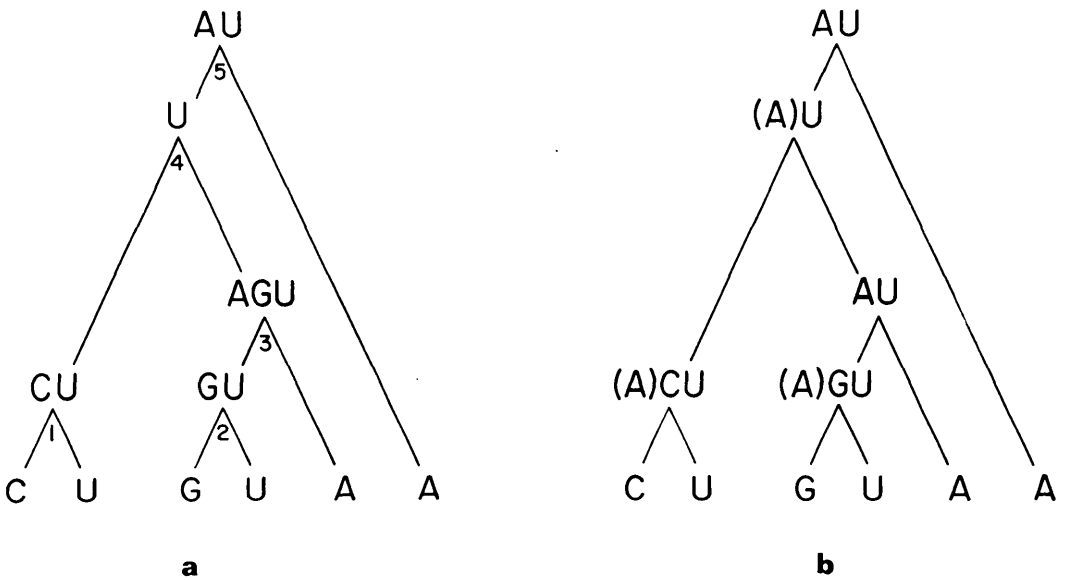

FIG. 1.—Reconstruction of Ancestral Nucleotide Forms. On the left side, Figure 1a shows the preliminary reconstruction beginning with an assumed topology and an assignment of neucleotides to the present taxonomic units at the bottom. Each ancestral set is composed of those nucleotides common to its immediate descendants, if any, otherwise it is composed of all of them. Arabic numerals indicate the set number and their order of construction. Figure 1b on the right shows the result when the preliminary phase has been converted to the final phase by the method described in the text. Nucleotides in parentheses have been added.

Reconstruction of possible ancestral nucleotides—preliminary phase

The preliminary reconstruction has been previously described (Fitch, 1970). That reconstruction proceeds from the descendent character sets by formulating an ancestral character set for an immediate ancestor and working backward, one successive node at a time, until finally the nodal character set for the most distant ancestor has been formed. The formulation of each nodal set follows the following simple rule: The preliminary nodal set shall be comprised of all characters (nucleotides) common to both immediately descendent sets; if none are common to both, then the preliminary nodal set will be comprised of all characters found in either. In mathematical terms, the nodal set is the intersection of its immediately descendent sets if the intersection exists (i.e., is not empty) otherwise it is the union of those sets. An example is shown in Figure 1a (on the left). Because there are so many different nucleotides in only 6 taxonomic units, many of the descendants have no nucleotides in common, with the result that the first (lowest) three preliminary nodal sets are formed by unions. At the penultimate node [4], the intersection $CU \cap AGU = U$ and so U is the preliminary nodal set. The ultimate node is once again a union. The example is chosen for its completeness in representing possible problems rather than for its representativeness. With real amino acid sequences such as the cytochromes *c*, the vast majority of the nodal sets are simpler in that they are formed by intersections of identical elements rather than by unions as in this case. It should be noted that for every occasion that a union is required to form the preliminary nodal set, a mutation (nucleotide replacement) must have been fixed in this nucleotide position at some point during

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

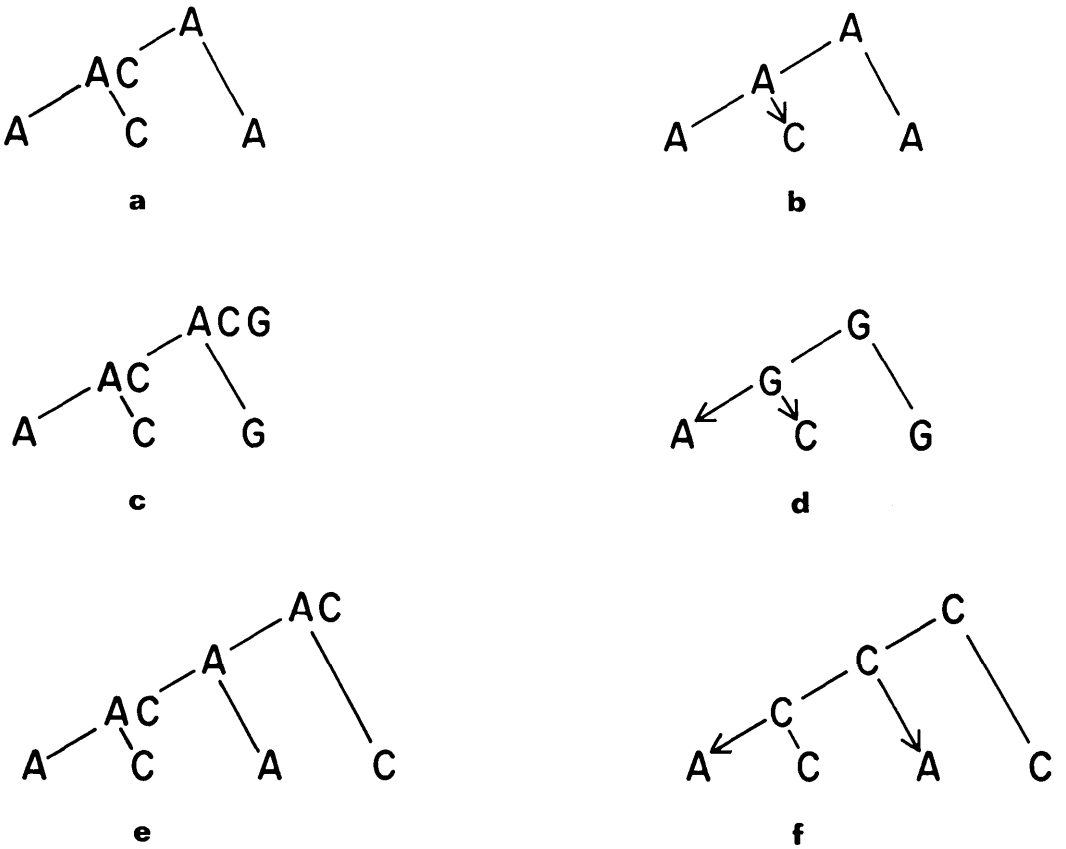

FIG. 2.—Deficiencies of Preliminary Phase Reconstruction. In each case, the trees on the left (2a top, 2c middle and 2e bottom) show the results of a preliminary phase reconstruction of ancestral nucleotides. In each instance, there is an interpretation, shown on the right, which is not inherent in the preliminary reconstruction. See text for explanation. Arrowheads are shown on segments to indicate the occurrence of nucleotide replacements.

the evolution of this position. Thus counting the number of unions gives one the minimum number of fixations (or changes of state) required to account for descendent nucleotides (characters) from a common ancestor, given the phylogeny assumed at the outset.

Reconstruction of possible ancestral nucleotides—final phase

To understand why there is a final phase following the preliminary phase, it is necessary to understand in what respects that preliminary phase may be deficient. These are shown in Figure 2.

In Figure 2a (upper left) is shown a preliminary phase reconstruction of a position from three taxonomic units. The ambiguous AC shown at the lower ancestral note represents an initial inability to decide whether the ancestral form was an A that was replaced by a C in the line of descent to the middle taxonomic unit or was a C that was replaced by an A in the leftmost line of descent. The only certainty is that a replacement is required. The third taxonomic unit, possessing an A, requires the ultimate ancestor to possess the A unless we postulate additional replacements beyond the minimum required to account for

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

the data. But this additional information about the upper ancestral node also makes it clear that the first node can not then be a C. The only formulation that will permit the descendent positions to be accounted for in a single replacement requires that replacement to be from A to C in the descent from the first node as shown in Figure 2b (upper right). The elimination of the C from the first node is determined by what may be called the *rule of diminished ambiguity*. Its precise formulation is encompassed in steps I and II of the algorithm, to be presented further on, that contains the complete set of rules for the final phase of reconstructing the nodal sets.

In Figure 2c (middle left) is shown another preliminary phase reconstruction which accounts, using two replacements, for the descent of the characters of the three taxonomic units given. Figure 3d (middle right), however, shows an equally adequate solution which is not encompassed by the possible alternatives available in Figure 3c. Clearly G is a valid alternative for the first node. This case is encompassed by the *rule of expanded ambiguity* which is precisely described in steps III and IV of the forthcoming algorithm.

In Figure 2e (lower left) is shown a third preliminary phase reconstruction that accounts for four descendants using two replacements. In Figure 2f (lower right) is an equally valid solution. Indeed, the C at the lowest node in the preliminary reconstruction is a valid alternative to the A if and only if a C is allowed at the penultimate node above. It is characteristic of this type of case that two nodes, separated by a single node, both contain a nucleotide not present in the intervening node because of the intersection process. Hence, this is called the *rule of encompassing ambiguity* which is formulated as step V of the forthcoming algorithm.

In the preliminary phase, the nodes in Figure 1 were formulated in the order of increasing ancestral remoteness (1→5, with the order for formulating nodes 1 and 2 being arbitrary). In the final phase, the order for correcting the nodal sets must be reversed (5→1).

The preliminary set for the ultimate node is made the final set for that node. We then go to the penultimate node (4 in this case) and proceed according to the following six step algorithm.

I. If the preliminary nodal set contains all of the nucleotides present in the final nodal set of its immediate ancestor, go to II, otherwise go to III.
II. Eliminate all nucleotides from the preliminary nodal set that are not present in the final nodal set of its immediate ancestor and go to VI.
III. If the preliminary nodal set was formed by a union of its descendent sets, go to IV, otherwise go to V.
IV. Add to the preliminary nodal set any nucleotides in the final set of its immediate ancestor that are not present in the preliminary nodal set and go to VI.
V. Add to the preliminary nodal set any nucleotides not already present provided that they are present in both the final set of the immediate ancestor and in at least one of the two immediately descendent preliminary sets and go to VI.
VI. The preliminary nodal set being examined is now final. Descend one node as long as any preliminary nodal sets remain and return to I above.

Figure 1 illustrates the operation of the algorithm. The left hand side (Figure 1a) depicts the preliminary nodal sets. The ultimate ancestral nodal set 5 (AU) is considered the final set and we turn our attention to preliminary nodal set 4. This nodal set does not contain an A and therefore, according to step I, we proceed to step III. Nodal set 4 was not formed by a union and therefore we are directed by step III to go to step V. Following the directions of step V we discover that A is present in both nodal sets 3 and 5 (the rule of encompassing ambiguity) and must therefore be added to nodal set 4. (Mathematically, $((1 \cap 5) \cup (3 \cap 5)) = AU$.

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

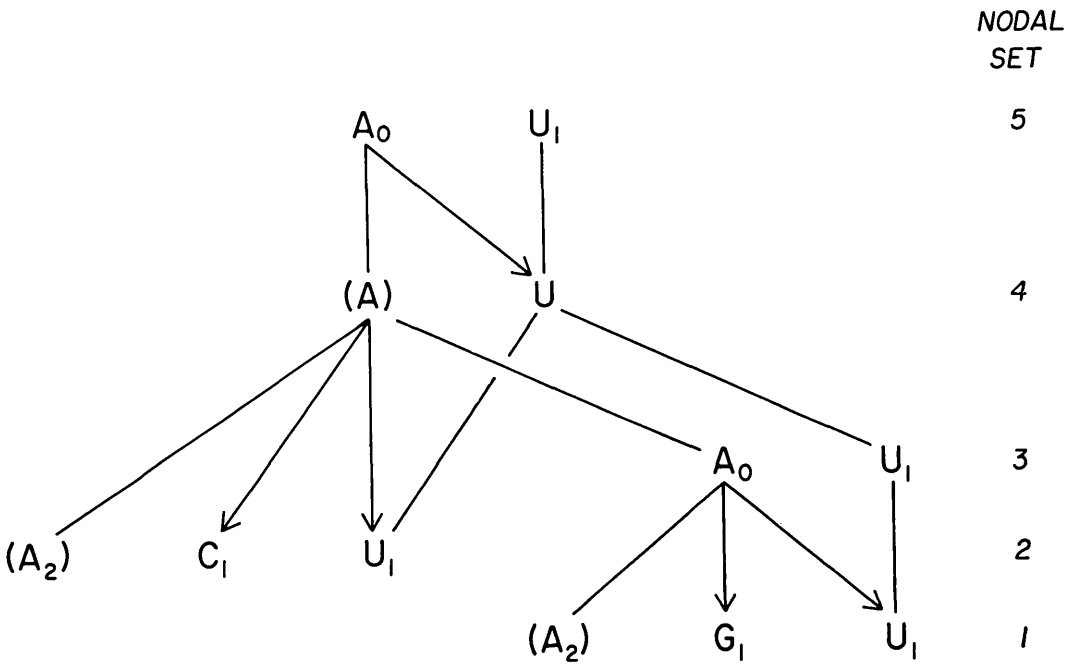

FIG. 3.—Possible Temporal Sequence of Ancestral Nucleotides. Nucleotides and topology identical to Figure 1 except only ancestral sets shown. Sequence is limited to follow the segments shown. Arrowheads denote nucleotide replacement. Where the original taxonomic unit(s) are direct descendants from one of the ancestral sets shown and where the nucleotide is different from the one indicated, a subscript has been added to indicate the necessary replacement. Thus on any given path of descent, arrowheads plus subscripts equal total replacements. See text for rules for finding the valid links between ancestral sets.

Since A is not present in nodal set 4, it must be added). The A is shown in parentheses to denote the fact that it was added in the final phase. We now proceed to step VI which tells us we are finished with node 4 and may proceed to node 3 and step I.

Preliminary nodal set 3, AGU, contains all of the nucleotides in its immediate ancestor and so we are directed by step I to go to step II. Step two tells us to eliminate all nucleotides in nodal set 3 not present in nodal set 4 (the rule of diminished ambiguity). Hence G is removed from nodal set 3 and we proceed to step VI which tells us nodal set 3 is now final and we may proceed to node 2 and step I.

Preliminary nodal set 2, GU, does not contain all of the nucleotides in its immediate ancestor and so we are directed by step I to go to step III. Because nodal set 2 was formed by a union, step III directs us to step IV. According to step IV, we must add any nucleotides in final nodal set 3 not present in preliminary nodal set 2 (rule of expanded ambiguity). Hence A is added and we proceed to step VI which tells us nodal set 2 is now final and we may proceed to node 1 and step I.

Preliminary nodal set 1, CU, like nodal set 2 before it, will pass through steps I, III, IV and VI, acquiring an A by the rule of expanded ambiguity. Since there are no further preliminary nodal sets remaining, the final phase of reconstruction is completed.

Permitted links between nucleotides in successive nodal sets

The ancestral forms have now been reconstructed such that all nucleotides that

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

could possibly occupy a given position in an ancestral sequence are shown. Any other nucleotides would require the postulation of additional replacements to account for final descendent forms. Even within the framework of the permitted nucleotides, not all replacements are allowable as one descends from node to node if one is to obtain a most parsimonious tree.

Figure 3 is a redrawing of the ancestral nodes of Figure 1b with links connecting those nucleotides that comprise valid lines of descent in a most parsimonious tree. The arrowheads are on those internodal links that involve a change of nucleotide state in descending from one node to the next. The subscripts give the number of nucleotide replacements required in those cases where one or both immediate descendants are not other ancestral nodes. Since node 4 has only other ancestral nodes for its immediate descendants, its nucleotides have no subscripts. The total number of replacements required must be four since there were unions used to construct nodes 1, 2, 3 and 5 as shown in Figure 1a. In Figure 3, a complete phylogeny will require a network composed of one nucleotide from every node using only the links shown. Such a complete linkage will involve nucleotides with subscripts and possibly links bearing arrowheads. The sum of the arrowheads plus subscripts on any complete valid linkage is four, the minimum number of replacements required to account for the data. There are 11 such linkages. One of them begins with the U in the ultimate ancestor, the other ten begin with the A. Of the ten beginning with A, one involves the replacement of A by U in the descent to node 4. The remaining are the 9 possible combinations of the three alternatives in descending to node 2 from an A in node 3 and the three alternatives in descending to node 1 from an A in node 4.

How then does one discover all possible valid linkages such as those shown in Figure 3? First of all, the nucleotides without parentheses were placed in the nodal set during the preliminary phase of reconstruction while those with parentheses were placed there during the final phase. Such nucleotides will be represented as N_i and (N_i) where the subscript denotes one of the four nucleotides. We shall use an arrow ($\rightarrow$) to denote descent from an ancestor to the next node. Thus $(N_i)\rightarrow N_j$ means that an ancestral nucleotide that was added to a nodal set during the final phase of reconstruction is replaced (since $i \neq j$) by a nucleotide that was originally assigned to the descendent nodal set in the preliminary phase. An example is shown by the $(A)\rightarrow C$ change in Figure 3.

The rules then are as follows: Given that a particular ancestral nucleotide is i,

I, $N_i \rightarrow N_i$ or $(N_i) \rightarrow N_i$ is obligatory if the descendent N_i exists; if the descendent N_i does not exist, then

II, All possible linkages are permitted except $N_i \rightarrow (N_j)$ and $(N_i) \rightarrow (N_j)$.

The interpretation of these rules may be seen in Figure 3. The first rule states that, given a particular ancestral nucleotide, there is no option but to link that nucleotide to the identical nucleotide in the descendent nodal set if that descendent nucleotide was placed there in the preliminary phase of the reconstruction. It is the operation of this rule that causes there to be only one valid complete linkage for Figure 3 when the ultimate ancestral form is a U. Another way of phrasing this rule is: A nucleotide not in parentheses is the only valid terminal point of a link which originates from the same nucleotide in its immediate ancestor.

When the first rule does not apply, then and only then does the second rule apply. As a consequence, except for the descent to the A at node 3, all ancestral A's are permitted to descend to any nucleotide in the nodal set of its immediate descendant.

None of the exceptions to rule II is shown in Figure 3. One may illustrate the exceptions though by imagining two cases. If the nodal sets were correct in Figure 3 except for the addition of a G to nodal set 4, that G could link to C or U in nodal set 1 but not to (A) because $N_i \rightarrow (N_j)$ is forbidden.

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

Alternatively, if the nodal sets were correct in Figure 3 except that the U in nodal set 1 were (U), then the link shown from the A in node 4 to the (now parenthetical) U in node 1 would have to be removed because $(N_i) \rightarrow (N_j)$ is forbidden. Another way of phrasing the exceptions to rule II is: links with arrowheads may not terminate at nucleotides in parentheses.

"Probabilities" associated with specific links

Although the total number of fixations ascribed to a single nucleotide position is the same for all of the most parsimonious trees, it should be clear from an examination of Figure 3 that which nucleotide replaces which is a function of the particular set of valid linkages that is examined. Indeed even the same nucleotide replacement may occur on different links depending upon the set of linkages examined. What then is the best estimate of the weight that should be assigned to any given link observed in the various, most-parsimonious trees for the time period represented?

A caveat is necessary here. For real world descendent sequences, there will be historical events of which there is no evidence in the data. Thus we may know that a descendent G was an A in its immediately ancestral node. We act (compute) as if the historical event were A→G but it might really have been A→C→G. Our weighting procedure restricts its consideration to those events for which evidence exists and we obtain "probabilities" which add up to one. But since our computational world is more circumscribed than the less parsimonious universe of all possible events that would account for the descendent nucleotides, every such probability necessarily overestimates the likelihood that that event actually occurred in the evolutionary history of that nucleotide position; hence the quotation marks. We pursue this computation in the face of this deficiency because 1, for many purposes we do not need to know the likelihood of events for which there is no evidence, and 2, the resultant probabilities are the most rational means of weighting alternative possibilities when tabulating what nucleotide replacements occurred how often in which positions and in which internodal intervals. Such tabulations are necessary in turn in order to answer questions about the randomness (or lack thereof) with which nucleotide replacements distribute themselves.

One might assume that all possible valid sets of complete linkages are equiprobable. Put another way, this assumes that every most parsimonious tree is as good as another. Behind this kind of statement is the false assumption that there is a randomness to the selection of taxonomic units and to their assignment to the branch tips of the three. In the present example shown in Figure 1, it leads to the absurd conclusion that the odds are 10 to 1 in favor of the ancestral nucleotide being an A. This is particularly absurd in view of the fact that had we not known about the right-most, distantly related taxonomic unit with its A, all ancestral nodal sets would have been U and only U in the most parsimonious solution. There is generally, in this manner of estimating the probable ancestral nucleotide, a bias in favor of the predominating nucleotide in the line descending from the ultimate ancestor that has the fewest bifurcations.

An alternative procedure assumes that all permitted nucleotides for the ultimate ancestor are equiprobable (i.e., in Figure 4, $P(^5A) = P(^5U) = 0.5$, where ^kN denotes a nucleotide in the k^{th} node). This alternative also assumes that every valid link from a given nodal nucleotide, $L(^kN)$, is equiprobable.[1] Thus, the probability that a given link is correct is $P[L(^kN)] = P(^kN)/n$ where n is the number of links descending from ^kN. Therefore, since there is only one link from ^5U, then $P[L(^5U)] = P(^5U) = 0.5$.

[1] This assumption requires that there be no bias with respect to which of three nucleotides replace the fourth. This is approximately true but there is evidence the G→A replacement occurs more frequently than would be expected if replacements were random (Fitch, 1967; Vogel, 1969).

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

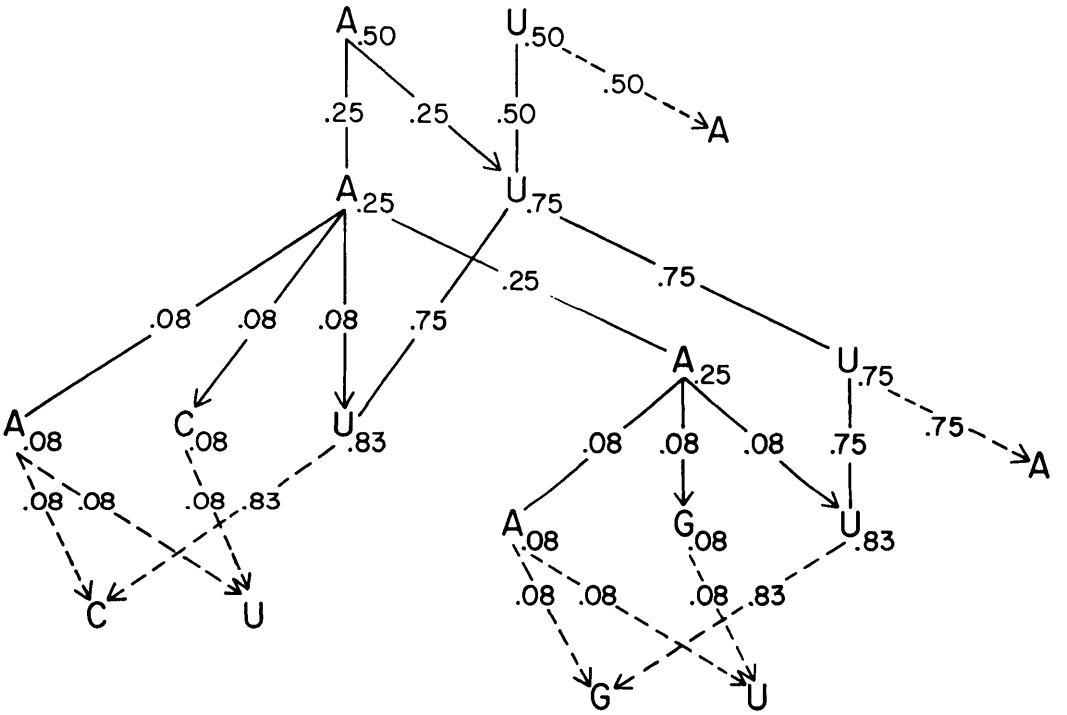

FIG. 4.—"Probability" That Segments are Part of the Temporal Sequence of Descent. Nucleotides and topology identical to Figure 3 except that the descendent taxonomic units have been added and connected to their immediate ancestral set by a dotted segment in those particular cases where a nucleotide replacement would be involved. Numbers indicate the probability that a given nucleotide or segment link would have been the one used under the assumption made in the text.

For the two links from ^{5}A, P[L(^{5}A)] = P(^{5}A)/2 = 0.25. The descendent nodal nucleotides have probabilities associated with the sum of the probabilities for the links that terminate upon them. Thus P(^{4}A) = 0.25 and P(^{4}U) = 0.25 + 0.5 = 0.75. Note that the sum of the probabilities for the nucleotides in any given nodal set must equal one and the same is true for the sum of the probabilities for the links between any two consecutive nodes. Figure 4 depicts the probabilities associated with every link and every nodal nucleotide shown in Figure 3. Also shown, by dotted lines, are the probabilities associated with those links to a final descendant where a nucleotide replacement was required. The result is that four necessary nucleotide replacements are distributed among those that form a part of a most parsimonious tree as follows: A→C, 2/3; A→G, 1/6; A→U, 7/12; C→U and G→U, 1/12; U→A, 3/4; U→C and U→G, 5/6.

The following relationships may be noted as indicating the "reasonableness" of this procedure. (I), C and G are each present in only a single taxonomic unit and are given only low probabilities of having existed at ancestral nodes. That is to say, where a phenotype is singular we would expect it to have arisen since the most recent bifurcation. (II) A, which has no representatives among the descendants of nodes 1 and 2 is given only a small probability of being the ancestral form at these nodes. The probability of the A, C and G nucleotides representing the ancestral form in nodes 1 and 2 for the case where all complete linkages are equiprobable would be more than three times as great as those shown in Figure 4.

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

DISCUSSION

A method for reconstructing ancestral sequences was first presented by Fitch and Margoliash (1967). That early method was also based upon the concept of minimum evolution but was less formal in that all possible solutions might not necessarily be recognized. Moreover, in order to decide among several equally parsimonious alternatives, the expected influences of selection were invoked. For example, Figure 2f shows one of three possible ways of accounting for the descent (on the topology given) of the four nucleotides presented. Clearly, this requires the same replacement, C→A, to have occurred twice. We may consider this pair of replacements to represent a parallel fixation. If, however, the two lower ancestral nodes are set equal to A rather than C while leaving the ultimate ancestral node at C, then we are indicating that a distant C→A replacement was subsequently followed by an A→C replacement. We may consider this pair to represent a back mutation (fixation). We believe that selection was more likely to discover the utility of a mutation in two closely related, and possibly contemporaneous, lines of descent than that a mutation once found beneficial should then become deleterious relative to its previous ancestral form. This remains our belief to the extent that selection is the operative mechanism. However, in view of the recent speculation about the possibility of neutral mutations (Kimura, 1968; Smith, 1968; Arnheim and Taylor, 1969; King and Jukes, 1969; O'Donald, 1969; Corbin and Uzzell, 1970), it is perhaps best not to exclude those back mutations for which parallel fixations represent an equally parsimonious solution. Actually, the procedure in this paper produces weights suggesting that back mutations are less likely on statistical grounds since the probability of the C→A replacement in the descent to the penultimate ancestor in Figure 2f is only 0.25.

Finally, it should be pointed out that the procedure presented here has not been mathematically proved. Counter examples have been diligently sought but all attempts to find a more parsimonious solution to the one(s) obtained by this method or to find equally parsimonious solutions not obtained by this procedure have failed. Nevertheless, in the absence of rigorous proof of the method's validity, a modicum of judicious reserve is not totally unwarranted. The procedure is presented in this state in order to invite mathematicians to bring their talents to bear upon this problem. One additional caveat is in order when translating amino acid sequences into codon sequences, namely, a selection must be made between the codons A.G.(CU) and U.C.(ACGU) for serine, A.G.(AG) and C.G.(ACGU) for arginine, and U.U.(AG) and C.U.(ACGU) for leucine. Any attempt at total ambiquity leads to such cases as (AU).(CG).(ACGU) for serine. Unfortunately, this implies the possibility that A.C.A. (threonine), U.G.U. (cysteine) or others are present when in fact only serine is present. This in turn could lead to others errors. A computer procedure is available for translating amino acids into codons that selects the codons for serine, arginine and leucine that are most likely to give the fewest mutations.

ACKNOWLEDGMENTS

This project received support from NSF grant (GB-7486). The University of Wisconsin Computing Center, whose facilities were employed, also received support from NSF and other U. S. government agencies.

REFERENCES

ARNHEIM, N., AND C. E. TAYLOR. 1969. Non-Darwinian evolution: Consequences for neutral allelic variation. Nature, 223:900–903.

CAMIN, J. H., AND R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution, 19:311–327.

CORBIN, K. W., AND T. UZZELL. 1970. Natural selection and mutation rates in mammals. Amer. Naturalist, 104:37–53.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool., 18:374–385.

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014

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

FITCH, W. M. 1969. Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. J. Mol. Biol., 26:499–507.

FITCH, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113.

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

KIMURA, M. 1968. Evolutionary rate at the molecular level. Nature, 217:624–626.

KING, J. L., AND T. H. JUKES. 1969. Non-Darwinian evolution. Science, 164:788–799.

O'DONALD, P. 1969. "Haldane's dilemma" and the rate of natural selection. Nature, 221:815–816.

SMITH, J. M. 1968. "Haldane's dilemma" and the rate of evolution. Nature, 219:1114–1116.

SOKAL, R. R., AND P. H. A. SNEATH. 1963. The principles of numerical taxonomy. Freeman, San Francisco. 359 pp.

VOGEL, F. 1969. Point mutations and human hemoglobin variants. Humangenetik, 8:1–26.

WAGNER, W. H., JR. 1961. Problems in the classification of ferns, p. 841–844. *In* Recent Advances in Botany. University of Toronto Press, Toronto.

WAGNER, W. H., JR. 1969. The construction of a classification, p. 67–99. *In* Systematic Biology, the Proceedings of an International Conference held at Ann Arbor in June 1967, publication 1692 of the National Academy of Science.

(Received October 12, 1970)

Downloaded from http://sysbio.oxfordjournals.org/ at University of York on August 17, 2014