TAXON 24(5/6): 609-613. NOVEMBER 1975

A SIMPLE TEST FOR THE POSSIBLE SIMULTANEOUS EVOLUTIONARY DIVERGENCE OF TWO AMINO ACID POSITIONS

*G. F. Estabrook and L. Landrum**

Summary

A taxonomic character for a group of organisms under study is said to be divergent or uniquely derived for that group if, during the evolution of that group, each character state arose only once. Homologous amino acid positions in homologous proteins can be thought of as taxonomic characters whose states are amino acids. In a similar way, DNA nucleotide positions can also be considered taxonomic characters. In the exposition below, we describe a simple test to determine when it is not possible for two amino acid (or nucleotide) positions to be divergent at the same time.

Several workers (Dayhoff and Eck 1969; Fitch and Margoliash 1967; Boulter *et al.* 1972) have made estimates, based on amino acid sequence data, of evolutionary relationships among taxa. The theoretical, as distinct from the technical, problems that attend the construction from amino acid sequence data of such estimates, are similar to the theoretical problems that challenge any taxonomist. Among these problems are those of homology, rate and divergence (Estabrook 1972).

In order to meaningfully compare protein A in taxon X with protein B in taxon Y we need some evidence or argument to support the claim that there is a taxon Z, ancestor of both X and Y, that contains a protein C from which A and B each evolved. Finding such evidence or making such arguments helps to solve the problem of homology. Needleman and Wunsch 1970; Sankoff 1972; and Wong *et al.* 1973 have recently suggested operational procedures for supporting or refuting statements of the form, "Protein A from taxon X and protein B from taxon Y are homologous".

The problem of rate has been discussed by several authors (Kirsch 1969; Sarich and Wilson 1966; Colless 1970). There are, in actuality, several problems of rate, for this concept implicates itself in a wide variety of factors that would militate against the credibility of objectively constructed estimates of evolutionary relationships. There are distinct concepts of evolutionary rate whose several definitions vary according to 1.) what is evolving: number of taxa; overall phenetic differences among taxa; patristic differences among taxa; the value of some measurable character of the taxa; and others, and according to 2.) the concept of quantification: counts; differences measures; mm/10^6 yr; probability of a unit change; and others. No matter how the concept of evolutionary rate is defined, the very attractive simplifying assumption that evolutionary rates are equal through time, or across phyletic lines is almost never true. How to define an effective concept of evolutionary rate that admits of a credible estimate of its value is also among the problems of rate. Fitch (1971) has recently given us an excellent example of how to work creatively and effectively with concepts of evolutionary rate.

The problem of defining and recognising divergence is the primary concern of the present discussion. Suppose there is a collection S of taxa whose evolutionary

* Dept. of Botany, University of Michigan, Ann Arbor, Michigan 48104.

relationships it is our purpose to estimate. Suppose further that there is a collection of apparently homologous proteins, one from each taxon in S, and for each of these proteins the amino acid sequence is known. An additional homology-like problem must now be solved: "What amino acid position in the protein from taxon X should be compared with what amino acid position in the protein from taxon Y?". Suggestions for solving this alignment problem have recently been made by Reichert (In Prep.). Suppose, lastly, that some alignment of the amino acid sequences of the homologous proteins in the taxa from S can be recognized as apparently correct (Sankoff 1972).

If taxon Z were the immediate ancestor of taxon X and if Z had Glu in position 74 while X had Asp in position 74, then, apparently, Asp was derived from Glu during the evolution of X from Z. It is possible for the taxa in S to have evolved from their most recent common ancestor in such a way that, to every amino acid to appear in a given position in the amino acid sequence (e.g. position 74), there corresponds at most one descendent taxon that did have that amino acid in that position but whose immediate ancestor did not. A given amino acid position that realizes this possibility is a divergent position. An amino acid position can also be construed as a taxonomic character whose states are amino acids. A divergent position would correspond to a divergent character.

An amino acid position in which the amino acid never changes during the evolution of the taxa in S would be divergent, but in a trivial sense that contributes little to revealing evolutionary relationships. However, sure knowledge of the divergence of a position in which several different amino acids have occurred would contribute greatly to the illucidation of evolutionary relationships, for a vast number of estimates of relative recency of common ancestry could be recognized as false. For example, suppose position 74 were known to be divergent, then the estimate of evolutionary relationships shown in Figure 1 can be recognized as false.

Fig. 1. The amino acid in position 74 of taxa W X Y and Z is shown in parentheses. These taxa could be the cotton, buckwheat, mung bean, and spinach respectively of Boulter *et al.* 1972.

Sure knowledge that two given positions were each divergent could limit possibilities even further, especially if a variety of amino acids occurred in each. A relatively few divergent and variable positions can make a relatively clear picture of the branching patterns of the phyletic lines.

Sure knowledge of divergence is not something that we are likely to enjoy in the near future, unfortunately. Presently, the concept of divergent position is only an ideal for no operational procedures exist for distinguishing divergent from non divergent (parallel or homoplastic) positions (Estabrook *et al.* 1975). In spite of what might seem to be a hopeless case, Le Quesne (1969) offers an ingenious, logical procedure in his uniquely derived character concept for making some progress toward understanding which taxonomic characters might be divergent. Here we generalize Le Quesne's idea to concepts that are applicable to protein sequence data, and offer a simple procedure for the analyses of these data.

For any given collection S of taxa whose cladistic history is to be estimated, there is an enormous but finite collection η of phylogenetic trees any of which could be postulated. For N taxa in S the number of bifurcating trees that could be postulated is well known to be the product of the first N-1 odd integers. The number of trees in η is, therefore, even larger. It remains, however, always finite.

Consider an amino acid position. Suppose that position were divergent. Then η could be partitioned into two subcollections: those trees that are possible, and those trees that are not possible. Thus, to every amino acid position (denote it with K), there corresponds a partition of η into two subcollections. Suppose a given tree in η were known to be true. Then we could easily determine whether a given position, K, could be divergent or not. In this way we can partition η into the same two subcollections: those trees whose truth permits the conclusion that the given position, K, could be divergent, and those trees whose truth permits the conclusion that the given position, K, could not be divergent. Denote the first of these two subcollections with [K]. Thus the second can be denoted with η–[K], and represents those trees known to be false were K known to be divergent. If K is an invariant position, then [K] = η, but [K] may be equal to η under other conditions as well. In particular, [K] = η if and only if there is only one amino acid that appears in position K of more than one taxon in S. For example, position 12 of Boulter *et al.*, (1972) contains Lys for all taxa except spinach that contains Asp, and sunflower that contains Thr. Position 12 could be divergent on any tree and thus [12] = η.

Consider another amino acid position L and its associated subset [L] of phylogenetic trees. If both K and L are divergent, the collection of possible evolutionary trees comprises those common to both [K] and [L] (i.e. the intersection of [K] and [L], [K] $\cap$ [L]. One of the following relationships always holds:

i [K] = [L]
iii [K] $\subset$ [L]
ii [K] $\supset$ [L]
iv [K] $\neq$ [K] $\cap$ [L] $\neq$ [L]

In case i, the partition of S into subcollections of taxa each containing the same amino acid in position K is the same as the corresponding partition of S for position L. For purposes of estimating evolutionary relationships in the present context, K and L may be considered the same if case i holds.

In case ii, the partition of S that corresponds to K is a hierarchical refinement of the partition of S that corresponds to L. Case iii is the same with the roles of K and L reversed.

In case iv, the assumption that K and L are both divergent limits the collection of possible trees beyond either of [K] or [L] considered individually. This is, perhaps, the most interesting and useful case. If either [K] or [L] is the same as η, then case iv always fails, and the development to follow becomes uninterestingly trivial.

Suppose (case iv) that [K] $\cap$ [L] is empty, i.e., [K] and [L] have no evolutionary trees in common. The assumption that positions K and L are both divergent has now restricted possible evolutionary trees right out of existence. If we wish to preserve the idea that one possible evolutionary tree in η is *the true tree* whose identity it is our objective to reveal, then we must conclude from our knowledge that [K] $\cap$ [L] is empty, that at least one, if not both, of K and L are not divergent positions. This is a generalized form that is applicable to protein sequence positions of Le Quesne's procedure for learning something about the logical possibilities that exist for divergent characters.

It is important to realize that the concepts η, K and [K] are not just ideal, as is the concept divergent, but perfectly operational: η is finite, the amino acids that occupy position K in the study collection S of taxa are known, and thus [K] can be determined. Thus we can *know* whether [K] $\cap$ [L] is empty or not for any given pair K, L of amino acid positions. However, these arguments teach us only that it is *possible* to know. To follow is a simple procedure that instructs us *how* to know.

List all the amino acids that occur in position K of any of the taxa in S across the top of a rectangular matrix. List all the amino acids that occur in position L of any of the taxa in S down the side of the same rectangular matrix. Place a mark in every box in the matrix for which there exists a taxon in S with 1.) the amino acid in position K, that labels the column the box is in, and 2.) the amino acid in position L, that labels the row the box is in. Tables 1 and 2 illustrate the procedure.

Start in any marked box and go from marked box to marked box subject to the following constraints:

Table 1

	Asp	Gln	Ala	Ser
Lys	x		x	
Asp				x
Asn	x	x	x	
Gln				x

K = position 4,
L = position 70
Boulter *et al.* 1972

Table 2

	Glu	Gly	Ser
Lys	x		
Asp		x	x
Asn	x	x	
Gln		x	

K = position 68,
L = position 70
Boulter *et al.* 1972

1. go next to a marked box that is either a) in the same row, or b) in the same column as the box you are currently in;
2. never go back to the box from which you have just come.

If no matter what box you start in, you cannot return to that box (i.e. it is not possible, given the pattern of marked boxes), then [K] ∩ [L] is not void, and it is possible for K and L to each be divergent positions. This case is illustrated in Table 2. If you can start in a marked box to which you can subsequently return, then [K] ∩ [L] is void, and it is not possible for both K and L to be divergent positions. This case is illustrated in Table 1.

A change in the amino acid in a given position can be construed as a change in the nucleotides comprising the codon that determines the amino acid for that position. Thus, the same concepts and procedures can be applied at the level of the sequences of nucleotides constituting the codons that transcribe for the amino acids that occupy positions K and L in the preceding discussion. In this application, one would speak of divergent nucleotide positions, and the matrices used in the test would be at most 4 boxes square, with the rows and columns corresponding to TACG or UACG. Because of the redundancy in the genetic code, in frequent cases only the first two of the three nucleotide positions constituting a codon can be used in this analysis. Examples of the procedure at the nucleotide level are shown in Tables 3 and 4. The two nucleotide

Table 3

	A	U	G
G	x	x	x
C		x	
U	x		

K = Position 63,
Nucleotide 2
L = Position 4
Nucleotide 1
Boulter *et al.* 1972

Table 4

	A	G	C
G	x	x	x
U	x		x
C	x		

K = Position 66,
Nucleotide 1
L = Position 4
Nucleotide 1
Boulter *et al.* 1972

positions compared in Table 3 can both be divergent, while the two nucleotide positions of Table 4 cannot both be divergent.[1]

Camin and Sokal (1965) called two cladistic characters *compatible* if they could both be divergent, and *incompatible* if it were logically impossible for both to be divergent. Estabrook *et al.* (in press) have proven that if every pair of cladistic characters in some collection is compatible, then it is possible for every cladistic character in that collection to be divergent. Unfortunately, this result is not valid for amino acid, or nucleotide, positions, in the sense that it is possible to have three positions that are pairwise compatible, and yet the positions cannot all three be divergent.

While these procedures do not teach us for sure which positions are divergent, it is clear that some positions change too slowly to reveal evolutionary relationships (e.g. when $\eta = [K]$, while others may change so rapidly that they will exhibit parallelisms and reversals that cloud the true picture of evolutionary relationships we seek to clarify. Those positions that admit of the *possibility* of being divergent when compared with each other, given a dearth of contrary evidence of other kinds, might be considered the most likely to be divergent. We might wish to consider first those hypotheses of evolutionary relationships that remain possible when we assume that those positions most likely to be divergent, actually are.

References

BOULTER, D., J. A. M. RAMSHAW, E. W. THOMPSON, M. RICHARDSON, and R. H. BROWN 1972 – A phylogeny of higher plants based on the amino acid sequences of cytochrome *c* and its biological implications. Proc. R. Soc. Lond. B. 181: 441-455.

CAMIN, J. H. and R. R. SOKAL 1965 – A method for deducing branching sequences in phylogeny. Evolution 19: 311-26.

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

DAYHOFF, M. O., and R. V. ECK 1969 – Inferences from protein sequence studies in Atlas of Protein Sequence and Structure, ed. M. O. Dayhoff, Silver Spring, Md.: Nat. Biomed. Res. Found.

ESTABROOK, G. F. 1972 – Cladistic methodology: A discussion of the theoretical basis for the induction of evolutionary history. Ann. Rev. Ecol. Syst. 3: 427-456.

ESTABROOK, G. F., C. S. JOHNSON and F. R. MCMORRIS – An idealized concept of the true cladistic character. Math. Biosciences. 23: 263-272.

ESTABROOK, G. F., C. S. JOHNSON, and F. R. MCMORRIS – An algebraic analysis of cladistic characters. Submitted to Discrete Math.

FITCH, W. M. 1971 – Rate of change of concomitantly variable codons. J. Molec. Evol. 1: 84-96.

FITCH, W. M. and E. MARGOLIASH 1967 – Construction of phylogenetic trees. Science 155: 279-284.

KIRSCH, J. A. W. 1969 – Serological data and phylogenetic inference: the problem of rates of change. Syst. Zool. 18: 296-311.

LE QUESNE, W. J. 1969 – A method of selection of characters in numerical taxonomy. Syst. Zool. 18: 201-205.

NEEDLEMAN, S. B. and C. D. WUNSCH 1970 – A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: 443-453.

SANKOFF, D. 1972 – Matching sequences under deletion/insertion constrains. Proc. Nat. Acad. Sci. 69: 4-6.

SARICH, V. M. and WILSON, A. C. 1966 – Quantitative immunochemistry and the evolution of primate albumins: micro-complement fixation. Science 158: 1200-1203.

WONG, A. K. C., T. A. REICHERT, D. N. COHEN, and B. O. AYGUN 1973 – A generalized method for matching informational macromolecular code sequences. Comput. Biol. Med. 3: 1-15.

1. W. M. Fitch presented a similar idea at the Classification Society Meetings 1974 Ann Arbor, Mich.