THEORETICAL FOUNDATIONS OF PALEOGENETICS

RICHARD HOLMQUIST
SPACE SCIENCES LABORATORY
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction

This paper presents mathematical reflections that were completed during the Spring of 1966 (Holmquist [9]). In light of increasing knowledge and interest in the evolutionary significance of the primary amino acid or nucleotide base sequences of homologous (see Section 3.2 for definitions) proteins or DNA's, both within and among various phylogenetic species, these calculations are given here in the hope that they may prove useful to a wider audience. The mathematics, though straightforward, is complex. Therefore, a conscientious effort has been made to relate the mathematical equations to concrete physical phenomena so that the paper may be more readable to mathematicians for whom the historical jargon of molecular biology may be unfamiliar as well as to the biologists, biochemists, and anthropologists who may want to use the mathematics as a tool for interpreting their experimental data.

The problems in molecular evolution that are soluble by a study of protein sequences and homologies may for convenience be divided into three classes: (a) the construction of phylogenetic trees; (b) the deduction of the primary amino acid sequences of the common ancestral proteins at the branch points of the phylogenetic tree; and (c) the assignment of a time scale to each leg of the phylogenetic tree. Historically, three concepts have been extremely useful in solving these problems: amino acid differences between homologous proteins (Zuckerkandl and Pauling [28]; Needleman and Wunsch [18]), the minimum mutation distance between homologous proteins (Jukes [10]; Fitch and Margoliash [4], [5]), and the path of least information or maximum entropy (Reichert and Wong [21]) between two homologous proteins.

It has seemed plausible to attempt to correlate the amino acid or base differences between two homologous proteins or nucleic acids with a time of origin, measured from the present, of a "common ancestor" protein or nucleic acid which is homologous to both. In favorable cases a possible primary sequence for the common ancestral molecule can be deduced. Proceeding in this way, one can build up a biochemical tree of life which can be compared to those

The preparation of this manuscript was supported by NASA Grant NgR 05-003-020, "The Chemistry of Living Systems."

evolutionary and phylogenetic relationships that are already known from classical biology. The name *paleogenetics* or *paleobiochemistry* has been suggested for that branch of science which concerns itself with molecular restoration studies of the above kind (Zuckerkandl and Pauling [28]; Pauling and Zuckerkandl [20]; Zuckerkandl [27]). Among the proteins for which such studies have been made, one may mention the hemoglobins of a great many species (see above references), the cytochromes *c* of various organisms (Margoliash [15]), the A and B fibrinopeptides (Doolittle and Blombaeck [2]), and the ferredoxins (Matsubara, Jukes, and Cantor [17]). In each case, the primary amino acid sequences of the proteins were used as base data. Wilson and Sarich [26] have used quantitative immunological techniques to study albumin and transferrin. Paleogenetic studies have been made on deoxyribonucleic acids by Martin and Hoyer [16] and by Kohne [14]. These investigators used hybridization, kinetics of renaturation, and thermal denaturation techniques to establish the degree of similarity of various mammalian DNA's and the evolutionary relationships between them. A more complete tabulation of primary sequence data may be found in Dayhoff's *Atlas of Protein Sequence and Structure* [1].

The success of the above methods is attested to by the general concordance of their results with the paleontological fossil record.

Infrequently, the evolutionary relationships revealed by studies of the above type differ radically (Doolittle and Blombaeck [2]; Wilson and Sarich [26]) from those that have been deduced from a great body of classical biological and paleontological evidence. The resolution of such differences remains a current problem.

Nevertheless, with the exception of the paper by Reichert and Wong, theoretical, quantitative, justification of the computational methods employed with each of these approaches has been lacking. It has been no more than fortuitous that for most of the proteins which have been studied to date, the mutation rate has been relatively low. The situation is particularly confusing when the mutation rate is moderate, for then the low rate approximations are being used at or above the limits of their validity; however, the protein sequences still haven't been completely randomized, so that evolutionary information is still present. The problem is "how much," and how to extract it, with confidence in the final result.

In order to interpret the experimental data, which consists of (in decreasing order of information content) the comparison of the primary nucleotide sequences, the primary amino acid sequences, the number and kind of amino acids, the amino acid compositions, the number and kind of nucleotide bases, and the nucleotide base compositions, a more sound theoretical foundation is needed. Two contributions to this theory may be found in the papers by Jerzy Neyman [19] and by Reichert and Wong [21]. Neyman particularly clearly defines, and emphasizes the complexity of, the statistical, topological, and temporal aspects of paleogenetics; and he makes a mathematically rigorous quantitative beginning towards solving them. Reichert and Wong approach

these problems from the viewpoint of set theoretic, informational, and thermodynamic principles. Fitch and Margoliash [4] have utilized the concept of "minimum mutation distance" to construct phylogenetic trees. Although this parameter is not mathematically the "best," and gives incorrect results when the mutation rate is high (it ignores, for example, multiple hits at the same nucleotide site and back mutations, and requires not generally correct *a priori* assumptions about the time sequence of mutational events), it has been extremely useful in making sense out of the mass of data now available. Gatlin [6] has suggested that living systems "may utilize the principle of Shannon's Second Theorem (Shannon and Weaver [23]) which states that it is possible to reduce transmission error without undue sacrifice of message variety or rate by properly encoding the message. These coding devices may form a quantitative basis for evolution and differentiation." King and Jukes [13] have pointed out the fact that evolution is fundamentally non-Darwinian in character: "natural selection is the editor, rather than the composer, of the genetic message."

There exist two theoretical approaches for relating the observed number of nucleotide base or amino acid differences between two homologous nucleic acids or proteins: minimizing the energy required to effect these changes and maximizing the entropy change. The concept of "minimum mutation distance," mentioned above, belongs to the first class. However, the energies required to interconvert one base to another are all of the same general order of magnitude, and it is therefore entropic factors that are usually the determinants of these differences. Also, the "minimum mutation distance" is sometimes a gross underestimate of the true number of primary mutagenic events that have occurred, is related in no simple way to these events, and has no firm theoretical foundation, except that it does state a minimum below which the number of primary mutagenic events cannot go. Although it may be useful in establishing the approximate topology of a phylogenetic tree, it does not suffice, particularly when the mutation rate is relatively high, to accurately establish the length of the legs of these trees: accurate values of these lengths are an absolute necessity if macromolecules are to be used as evolutionary clocks. In the latter respect, two summers ago I had the privilege of visiting the Olduvai gorge. Considering the landscape and the vastness of the African continent, one cannot help but admire the monumental and scientifically productive efforts of Dr. and Mrs. Leakey and their colleagues. As molecular evolutionists we do well to remember that the absolute time scales that we use come directly from the fossil record and not from macromolecules.

Whether the "ticks" of the evolutionary clock be electromagnetic, thermal, chemical, or other, in origin, the molecular quantity most closely and simply related to these "ticks" is the number of one step nucleotide base changes within DNA. At the observational level, these primary events are reduced by multiple hits at the same base site, back mutation, and the chance coincidence of having the same base at a given site in two homologous nucleic acids. At the protein level additional correction must be made for multiple hits within the same

codon, codon degeneracy, and the possibility that two homologous sites will have the same amino acid there by chance. The purpose of this paper is to show how to make corrections for these phenomena accurately.

The methods presented here differ in three important aspects from earlier approaches. First, they are more general: the only starting assumption is that the mutagenic events occur randomly along the polynucleotide sequence which codes for a particular protein. The mutation rate may have any numerical value and any temporal dependence; the nucleic acid segment, or the protein for which it codes, may be of any length. If, in fact, the mutagenic events occur nonrandomly along the polynucleotide sequence, it is inherent in the method to detect such nonrandomness. Second, they are more exact: the phenomena of multiple hits at the same nucleotide site, back mutation, and accidental identity at the same site are quantitatively accounted for without approximation. The effect of amino acid codon degeneracy and of multiple nucleotide base changes at different base sites within the codon triplet are evaluated. The formulas developed herein thus include as special cases those methods based on a low mutation rate. And third, the probability of obtaining *exactly* a given value of a parameter of interest is calculated so that not only the average and most probable values of these parameters are known, but also their variances. This latter fact establishes quantitative objective criteria for the significance of any computed or observed values of those parameters.

Before we proceed to the mathematical theory and derivations, the results of these calculations and their significance for paleogenetics will be summarized and discussed in the next section so that the more important points do not become lost in the necessarily lengthy mathematical development. Following these results, the abstract formulas from which they were obtained will be derived; and finally, an Appendix illustrating the formulas by numerical example, based on actual experimental data, will conclude the paper, so that others may be able to do similar calculations themselves on experimental data of their own choosing.

2. Results

Let us examine the first row in Table I. The first column shows that we are considering, for illustration, a DNA segment of $L = 18$ base residues which codes for a hexapeptide (second column $T = 6$). This DNA segment evolves to two present day homologous DNA's. The number of one step base changes or hits (See Section 3.2 for definitions) which separates each of these contemporary DNA's from the ancestral DNA is 9 (third column), or in an alternative viewpoint, the number of one step base changes which separates the two contemporary DNA's from each other is 18 (fourth column). However, some of the base changes will occur at the same base site, so that the number of different base sites hit in each homologue is less than 9, namely, 7.24 (fifth column). Because a base site hit more than once may revert (back mutate) to the same

TABLE I

AMINO ACID DIFFERENCES

L = number of nucleotide base sites ($L = 3T$),
T = number of amino acid sites,
X = primary mutagenic events (hits),
$N(x)$ = average number of base sites hit at least once,
$N'(x)$ = average number of base sites different from corresponding sites in the ancestral DNA,
$N(D)$ = average number of base differences between two present day homologous DNA's,
$N(A)$ = average number of amino acid sites different from corresponding sites in ancestral fibrinopeptide fragment,
$N(d)$ = number of amino acid differences between two present day homologous fibrinopeptide fragments.

Numbers in parentheses are calculated for restricted mutation: purine to purine; pyrimidine to pyrimidine.

L	T	X	$2X$	$N(x)$	$N'(x)$	$N(D)$	$N(A)$	$N(d)$
18	6	9	18	7.24	6.74	10.12	3.96	5.14 ± 0.86
					(5.88)	(7.91)	(3.28)	(4.65 ± 1.02)
9	3	4.5	9	3.70	3.46	5.15	2.28	2.62 ± 1.51
					(3.04)	(4.03)	(2.12)	(2.40 ± 1.38)
18	6	2	4	1.94	1.94	3.61	1.33	2.35 ± 1.20
					(1.94)	(3.46)	(1.23)	(2.19 ± 1.18)
9	3	1	2	1	1	1.85	0.76	1.32 ± 0.86
					(1)	(1.78)	(0.65)	(1.16 ± 0.84)

base as in the ancestral DNA, the number of base sites in each homologue which differ from the homologous sites in the ancestral DNA is still less, or 6.74 (sixth column). The number of base differences between the two present day DNA's will be twice 6.74 less that number of homologous sites, which though differing from the ancestral site, are the same by chance coincidence. The net result is that the two contemporary DNA's will differ in 10.12 (seventh column) homologous sites, on the average. Now the number of amino acid differences, 3.96 (eighth column), between the ancestral hexapeptide coded for by the ancestral DNA, and either of the two present day homologous peptides will be less than 6.74, because some of the differing base sites may fall within the same codon triplet, and also because some amino acids are coded for by more than one triplet. The number of amino acid differences between the two contemporary homologous hexapeptide is twice 3.96 less that number due to chance identity, or a total of 5.14 differences (ninth column), on the average.

The other rows in Table I are interpreted similarly and are discussed more exhaustively in the analysis of the fibrinopeptide A sequences in Section 6.3.

The first row of Table I demonstrates that the expected number of amino acid differences between two homologues may be less than the number of primary mutagenic events by as much as a factor of 3.5. Table I also clearly brings out the fact that the number of amino acid differences is not only not proportional to the number of mutagenic events (we would not expect it to be because of multiple hits, revertants, hits within the same codon, code degeneracy, and

chance coincidences), but it is not even a function only of the proportion of sites hit, X/L: for in the first two rows the numbers of amino acid differences are in a ratio of about 2, whereas X/L is the same ($\frac{1}{2}$) in both cases. This non-proportionality is even more dramatically illustrated by the second and third rows where the ratio of the X/L's is 4.5:1, whereas the number of amino acid differences is almost identical.

Five points need emphasizing.

(1) Exact calculations put very definite quantitative limits on the permissible values of the mutation rate, time of divergence, and the number and distribution of hyper- or hypovariable sites.

(2) Methods based on the proportion of sites hit or unhit are inherently incapable of yielding physically meaningful calculations in some cases. This includes methods in which the data is "normalized" to 100 residues.

(3) The "$N(d)/2$" approximation (Zuckerkandl and Pauling [28]) for X is sometimes a poor one; for example, in the first row of Table I, $N(d)/2 = 2.07$, while the true value of X is 9.

(4) The "negative log" approximation (Zuckerkandl and Pauling [29]) for X may be quite inaccurate; for example, for the first row of Table I the true value of X is 9, while the negative log estimate is 5.74.

(5) The parameter X, or its minimum or maximum value consistent with the experimental data, should be used in constructing phylogenetic trees, not the "minimum mutation distance."

3. Derivations for DNA

3.1. *General.* Amino acid differences among proteins arise from mutational changes that occur in the nucleic acid segments which code for these proteins. The number of amino acids which have mutated in the proteins may differ from the number of mutations which have affected the nucleic acid segment for several reasons: (a) multiple hits at the same site—several mutagenic events occur at a single nucleotide position in the segment instead of each event occurring at a different nucleotide along the segment; (b) back mutation—a *multiply* hit single nucleotide site may end up as the same nucleotide as it was originally; (c) since each amino acid in a protein is coded by a triplet of nucleotides, several of the mutagenic events may fall within the same triplet—this would give rise to only a single amino acid substitution; (d) degeneracy—some amino acids are coded for by more than one nucleotide triplet so that a mutagenic event occurring within this triplet need not lead to an amino acid substitution; (e) viability—if a particular nucleotide mutation leads to a nonviable organism, one will not be able to observe this mutation as an altered nucleotide base or as an amino acid substitution; a mutagenic event that resulted in the formation of one of the three chain terminating (nonsense) codons might in some cases fall in this category; and finally, (f) even though the rate of mutation may be

accurately known over a certain region of space, what one must frequently examine is a particular subregion of this space; for example, if the mutation rate along a chromosome were known, the mutations themselves would show up as changes in the amino acid sequences of many nonhomologous proteins, only one, or even only a part of one of which is at hand for study. The *observed* mutations are therefore very much a function of the particular proteins or nucleic acids that are selected for analysis.

From the considerations of the preceding paragraph it is clear that no simple relationship exists between mutagenic events and observed protein mutations. As a consequence, a detailed understanding of paleogenetical studies requires the quantitative evaluation of each of the above factors so that their relative importance can be assessed.

3.2. *Definitions.* *Homologous* proteins are proteins that are *in fact* related to each other by point mutations in the common ancestral DNA coding for those proteins. It is possible for two proteins (or base sequences) to be identical or to bear any arbitrary degree of relatedness to each other without being homologous. This could occur by convergent evolution from two quite different ancestral DNA sequences. Such proteins are properly referred to as *analogous.* Experimentally it is usually difficult to distinguish between the two cases, but at times, when the mutagenic pathway is known, as with certain chemical mutagens, the distinction may be possible.

A *mutagenic event* is defined as a one step change of one nucleotide to a different nucleotide: that is, $C \rightarrow T = (C \rightarrow T)$, $C \rightarrow T$ $(C \rightarrow T \rightarrow C \rightarrow T)$, $C \rightarrow T$ (at one nucleotide site) and $A \rightarrow G$ (at another nucleotide site) represent, respectively one, three, and two mutagenic events, where C, T, A, and G are abbreviations for deoxycytidine, thymidine, deoxyadenosine, and deoxyguanosine, respectively. Such an event, of course, must be incorporated into the gene pool of the species if it is to be evolutionarily effective.

A nucleotide site is said to have been *hit* each time that a mutagenic event has occurred at that site.

A nucleotide site is said to have been *altered* if the nucleotide base occupying that site differs from the nucleotide base originally there before the mutagenic event occurred.

Other definitions will be introduced throughout the text as they are needed.

3.3. *Precise statement of problems to be solved in this paper.* (1) Consider a polynucleotide which contains L individual nucleotides. Let exactly X mutagenic events occur randomly along the length of this polynucleotide. After the X mutagenic events have occurred, in general, a number x, which is less than L, nucleotide sites will have been hit; for example, all X mutagenic events might occur at the same nucleotide site. Let $N(x)$ designate the average number of nucleotide sites which have been hit. An explicit formula for $N(x)$ is derived.

(2) The average number $N'(x)$ of nucleotide sites that have been altered will in general be less than $N(x)$ because of back mutations. An explicit expression for $N'(x)$ is given.

(3) An explicit formula for the average number of nucleotide base differences $N(D)$ between two homologous polynucleotides is derived, including correction for chance coincidences.

(4) Consider a protein of T amino acids which is coded by a polynucleotide of $L = 3T$ individual nucleotide bases. Let exactly X mutagenic events occur randomly along the length of this polynucleotide. After the X mutagenic events have occurred, a number A, less than T, amino acid sites will differ from the corresponding sites in the ancestral protein. An explicit formula for $N(A)$, the average number of amino acid substitutions that have occurred, is derived.

(5) Because of chance identities, the number of amino acid differences $N(d)$ between two homologous present day proteins will be less than $N_1(A)$ plus $N_2(A)$, where the subscripts refer to each homologue; a formula for $N(d)$ is derived.

(6) The limits of validity of the commonly used approximation $N(A) = N(d)/2$ are derived.

(7) Formulas are given which permit the proportion of amino acid substitutions which have occurred by one base, two base, and three base changes to be calculated.

3.4. *Calculation of $N(x)$: multiple hits.* Let us make the following definitions. An x part partition of X is a decomposition of X into a set of x (nonzero) positive integer summands $\{a_1, \cdots, a_x\}$, where $\sum_i a_i = X$. Partitions having the same a_i are considered to be identical even though the order of the a_i may differ in two such partitions. Let a particular x part partition of X be denoted by $(x, X)_j$, and let $n_{a_{ij}}(x)$ be the number of integers in this partition having the value a_i. Note that $\sum_{a_i \neq a_k} n_{a_{ij}}(x) = x$.

To make these abstract definitions more concrete, consider the following example. A particular 3 part partition of 6 is the set of integers $\{4, 1, 1\}$. Here, $a_1 = 4$, $a_2 = 1$, and $a_3 = 1$. We denote this particular partition by $(3, 6)_1$, where the subscript $j = 1$ is to remind us that this partition refers specifically to the set of integers $\{4, 1, 1\}$. For $(3, 6)_1 = \{4, 1, 1\}$, $n_{a_{11}} = 1$, $n_{a_{21}} = 2$, and $n_{a_{31}} = 2$. As stated in the definition, $a_1 + a_2 + a_3 = 4 + 1 + 1 = 6$; and $n_{a_{11}} + n_{a_{21}} = 3$. A different 3 part partition of 6 would be the set of integers $\{3, 2, 1\}$, and this partition could be labeled $(3, 6)_2$, for example.

Define N_{jx} as the number of ways of realizing $(x, X)_j$ along a polynucleotide which contains L individual nucleotides. This definition of N_{jx} requires that we associate the partitions $(x, X)_j$ in some well defined way with the polynucleotide of length L. We do this as follows. The partition (x, X) *means* that x nucleotide sites have been hit a total of X times; and the particular x part partition of X, $(x, X)_j = \{a_1, a_2, \cdots, a_x\}$, *means* that the first nucleotide site has been hit a total of a_1 times, the second site a_2 times, and the xth site a_x times. (The nucleotides in the polynucleotide can be numbered in any convenient manner: for example, the 3′ terminal nucleotide could be taken as the first nucleotide and the 5′ terminal nucleotide as the Lth nucleotide.) Now the first nucleotide site can be hit a_1 times in a number of ways: for example, if $X = 30$ and $a_1 = 3$, the first, second, and third mutagenic events could occur at the

first site, or alternatively the second, fifth, and twenty seventh mutagenic events could occur there. (The mutagenic events are numbered in any convenient manner.) Similar considerations hold for the other nucleotide sites. The total number of ways in which the first site can be hit a_1 times, the second site a_2 times, the xth site a_x times, and the $(x + 1)$th, $(x + 2)$th, . . . , and Lth sites zero times is *by definition* N_{jx}.

Now the average number $N(x)$ of polynucleotide sites that have been hit is by definition

$$N(x) = \sum_x xP(x), \tag{3.1}$$

where $P(x)$ is the probability that *exactly* x sites have been hit. But $P(x)$ is by definition

$$P(x) = \frac{\sum_j N_{jx}}{\sum_{x \leq L} \sum_j N_{jx}} = \frac{\sum_j N_{jx}}{L^X}. \tag{3.2}$$

The denominators in (3.2) are the total number of ways X mutagenic events can hit L nucleotide sites. Thus, if we can find an expression for N_{jx}, $N(x)$ will be given explicitly by (3.1). This expression for N_{jx} is derived in the following paragraph.

If x nucleotide sites have been hit in a polynucleotide of L nucleotides, then $L - x$ have not been hit. This can happen in

$$W_1 = \frac{L!}{(L - x)!x!} \tag{3.3}$$

ways. Now let us limit our consideration to those sites which have been hit at least once. In particular, let us assume these sites have been hit in the precise manner defined by the physical meaning attached to $(x, X)_j$. These x sites can be hit in a total of

$$W_2 = X!x! \tag{3.4}$$

ways, because $X!$ is the number of ways X mutagenic events can occur along the polynucleotide, and $x!$ is the number of ways that the x a_{ij} can be permuted among themselves. The factor $x!$ arises from the fact that in the *definition* of $(x, X)_j$ the order of the a_{ij} in the partition was irrelevant, while the *physical meaning* attached to $(x, X)_j$ was such that identical partitions in which the a_{ij} occur in different orders refer to different physical situations. Not all of these W_2 ways represent distinct physical situations, for the a_{ij} hits at the ith site can occur in $a_{ij}!$ ways and each of these ways leads to the same physical result. Similarly, if in the partition there are $n_{a_{ij}}$ integers having the value a_{ij}, these integers can be permuted among themselves in $n_{a_{ij}}!$ ways without altering the physical result. The total number of ways x sites can be hit by X mutagenic events is thus

$$W_3 = \frac{W_2}{\prod_{a_{ij} \neq a_{kl}} n_{a_{ij}}! \prod_{a_{ij}} a_{ij}!}. \tag{3.5}$$

Therefore,

$$N_{jx} = W_1 W_3 = \frac{X!}{\prod\limits_{a_{ij} \neq a_{kj}} n_{a_{ij}}! \prod\limits_{a_{ij}} a_{ij}!} \frac{L!}{(L - x)!}. \tag{3.6}$$

This completes the solution to our problem. *Tables of Factorials* $0! - 9999!$ [22] make it unnecessary to calculate the factorials in equation (3.6) by hand.

The method that has been given above for calculating $N(x)$ in terms of partitions illuminates the physical details of the mutation process. However, writing out the partitions that are needed in this method is frequently tedious. This is, if anything, an understatement. The basic difficulty is that no general formula exists for $n_{a_{ij}}$. In some applications, the calculation of the detailed structure for the probability for back mutation, for example, the individual a_{ij} and $n_{a_{ij}}$ of each partition must be known, and there is no way to get them except to write down the partitions one by one in some systematic manner that insures against leaving any partition out. In this respect, the *Tables of Partitions* [7] (see especially equation 1.1, p. ix, and equation 2.2a, p. xi) are very helpful, for they list the total number $P(X, x)$ of each possible (x, X) as well as the total number $p(X, X) = \sum_{x=1}^{X} P(X, x)$ of all possible partitions for a given X. Clearly, it would be desirable to have a formula for $N(x)$ that does not require the calculational labor of (3.6). Such a formula can be obtained as follows. Define $m(X, x)$ to be the number of ways X mutagenic events can hit x nucleotide sites, where each site is hit *at least* once. Mathematically, the number we have designated by $m(X, x)$ is the number of mappings of X *onto* x. Thus,

$$\sum_j N_{jx} = m(X, x) W_1, \tag{3.7}$$

where W_1 is given by (3.3).

We now calculate $m(X, x)$. The total number of ways X mutagenic events can hit x nucleotide sites is x^X. Therefore, $m(X, x)$ is given by x^X less those number of ways in which X mutagenic events can hit x nucleotide sites when k sites are not hit at all, where k takes on successively the values $0, 1, 2, \cdots, x - 1$. This follows from the fact that we *defined* $m(X, x)$ to include only those situations where *every* one of the x sites is hit *at least once.* Those situations in which some site or sites are not hit at all must be subtracted. But as in (3.3) the number of ways in which k sites can be hit and $x - k$ sites not hit is

$$W_4 = \frac{x!}{k!(x - k)!}. \tag{3.8}$$

Those k sites which have been hit can be hit in a total of $m(X, k)$ ways by definition. The total number of ways in which the X mutagenic events can hit x sites when some of the sites are not hit at all is thus

$$W_5 = \sum_{k=0}^{x-1} \frac{x!}{k!(x - k)!} m(X, k). \tag{3.9}$$

Therefore,

$$m(X, x) = x^X - W_5 = x^X - \sum_{k=0}^{x-1} \frac{x!}{k!(x-k)!} m(X, k). \tag{3.10}$$

Now $N(x)$ can be calculated from (3.10), (3.7), (3.2), and (3.1). It should be noted that because of (3.10) no knowledge whatsoever about partitions is required in calculating $N(x)$. Finally, we notice that (3.10) is a recursion formula for $m(X, x)$; that is, starting from $m(X, 0) = 0$, all the other $m(X, k)$ and finally $m(X, x)$ can be calculated from (3.10) alone. This important fact reduces the calculation of $N(x)$ to a simple iterative procedure which can readily be carried out by a computer. For those who are satisfied with the truth of (3.10) and are less interested in its physical derivation as given above, we comment here that it can be proved by mathematical induction on the positive integers x.

This completes the solution to the problem of multiple hits at the same nucleotide site:

$$N(x) = \sum_{x=1}^{L} xP(x), \tag{3.11}$$

$$P(x) = \frac{m(X, x)L!}{L^X x!(L-x)!}. \tag{3.12}$$

By maximizing $P(x)$ with respect to x, the *most probable* value of x can be determined. It is not obvious to the author whether this most probable value has an explicit mathematical formulation, or whether it must be determined by numerical calculation from (3.12).

3.5. *Calculation of $N'(x)$: back mutation.* After a_i mutagenic events have occurred at a given nucleotide site, the probability $P(a_i)$ that the final nucleotide is the same as the original nucleotide at that site depends only on a_i and on the number of nucleotides to which a given nucleotide can mutate. We give this probability for two models.

Case 1. Any nucleotide is free to mutate to any one of three other nucleotides.

Case 2. Any purine or pyrimidine nucleotide is free to mutate to only a purine or pyrimidine nucleotide, respectively.

Case 1 is given because amino acid substitutions are known which require the mutation of a purine to a pyrimidine or *vice versa.* Case 2 is given because of its possible usefulness with respect to chemical mutagens that are known to involve purine to purine or pyrimidine to pyrimidine transformations (C → U by nitrous acid, for example). As can be seen from Table II $P(a_i)$ is a *strong* function of the model selected and selection of the incorrect model in a particular case can completely invalidate the quantitative and qualitative topology of a paleogenetic analysis (such as the construction of a phylogenetic tree). Other models are possible, but the two given will suffice for present purposes.

The average probability of having the same nucleotide at any site after a_i mutagenic events have occurred there is

$$P = \sum_{a_i=1}^{X} P(a_i) \frac{\sum_{x \leqq L} \sum_{J} N_{jx} n_{a_{ij}}(x)}{\sum_{a_i=1}^{X} \sum_{x \leqq L} \sum_{j} N_{jx} n_{a_{ij}}(x)} \tag{3.13}$$

and the average number of altered nucleotides is

$$N'(x) = (1 - P)N(x). \tag{3.14}$$

TABLE II

$P(a_i)$ AS A FUNCTION OF a_i

The formula for $P(a_i)$ is proved by mathematical induction on the positive integers a_i.

	Case 1	Case 2
a_i	$P(a_i) = \frac{1}{4}\left[1 + \frac{(-1)^{a_i}}{3^{a_i-1}}\right]$	$P(a_i) =$ 1 if a_i even, 0 if a_i odd
0	1	1
1	0	0
2	1/3	1
3	2/9	0
4	7/27	1
5	20/81	0
.	.	.

Equation (3.13) appears more complicated than it really is (see Section A.2 of the Appendix, for a numerical calculation). However, its use does require a knowledge of the detailed structure of each partition $(x, X)_j$ for *all* possible x, and as stated on an earlier page, the only way to get this structure is by writing down *all* the partitions and counting the a_i in each to find the $n_{a_{ij}}$. If one is not interested in the *details* of the revertant process, then the net result, $N'(x)$, can be rapidly obtained for any X by repetitive application of the following recursion formulas which do not require a knowledge of the partition structure, and which may be proved by mathematical induction:

$$N'_0(x) = 0, \tag{3.15}$$

$$N'_{X+1}(x) = 1 + \left(1 - \frac{4}{3L}\right)N'_X(x). \tag{3.16}$$

Another very useful relation for reducing the calculational labor is given by

$$N'_{2X}(x) = 2N'_X(x) - \frac{4}{3}\frac{[N'_X(x)]^2}{L}. \tag{3.17}$$

Equations (3.15) through (3.17) are only valid for Case 1. For Case 2 the corresponding formulas are

$$N'_0(x) = 0, \tag{3.18}$$

$$N'_{X+1}(x) = 1 + \left(1 - \frac{2}{L}\right) N'_X(x), \tag{3.19}$$

$$N'_{2X}(x) = 2N'_X(x) - \frac{2[N'_X(x)]^2}{L}. \tag{3.20}$$

3.6. *Calculation of $N(D)$: nucleotide differences between homologous DNA's.* If the sequence of the common ancestor polynucleotide is known, then $N'(x)$ can be compared directly with the experimentally observed number of nucleotide substitutions for each homologue. In practice, the sequence of the ancestral polynucleotide is seldom known, and what one measures experimentally is the number and type of nucleotide differences $N(D)$ and coincidences $(L - N(D))$ between two homologous polynucleotides having the same common ancestor. A coincidence and a difference are defined, respectively, as the occurrence of identical or nonidentical nucleotide bases at the same position in the two homologous polynucleotides. The probabilities needed to calculate $N(D)$ are given in Table III.

TABLE III

PROBABILITIES NEEDED FOR CALCULATING $N(D)$

$p_i = [L - N'_i(x)]/L$, where the subscript i refers to homologue 1 or homologue 2.

	p_1p_2		$1 - p_1p_2$	
A	T	T	T	T
1	T	0	T	0
2	T	0	0	T
		$1 - p_1 - p_2 + p_1p_2$	$p_1(1 - p_2)$	$(1 - p_1)p_2$
		$1 - p_1 - p_2 + p_1p_2$		
1	0	G	G	0
2	0	G	0	G
Case 1:		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
Case 2:		1	0	0

The first column in this table indicates the ancestral polynucleotide A and the two homologues under consideration. The second column indicates the probability that both homologues have the same nucleotide base at a particular site as the ancestral polynucleotide. The last three columns indicate the probability that one or the other or both of the homologues has (have) a different nucleotide base at a particular site from the ancestral polynucleotide. For illustration, the ancestral base has been taken as thymidine. The last two columns clearly represent differences between the two homologues. Also, however, the third column, which represents those homologous bases that differ from the ancestral base at a given site, contains base pairs that may be the same (but not thymidine) or different. The probability for their being the same

is given in the third column of the lower part of the table (guanosine has been chosen for illustration), and the probability for their being different is given in the last two columns of the lower part of the table. Probabilities are given for both Case 1 and Case 2 (see Section 3.5). The total probability that the two homologues will differ from *each other* at any single site is therefore

$$\begin{aligned} p &= p_1(1 - p_2) + (1 - p_1)p_2 + (\tfrac{2}{3})(1 - p_1 - p_2 + p_1p_2) \\ &= (\tfrac{2}{3})(1 + \tfrac{1}{2}(p_1 + p_2) - 2p_1p_2) \end{aligned} \tag{3.21}$$

for Case 1, and

$$\begin{aligned} p &= p_1(1 - p_2) + (1 - p_1)p_2 \\ &= p_1 + p_2 - 2p_1p_2. \end{aligned} \tag{3.22}$$

for Case 2.

The probability $P(D)$ that exactly D differences will be observed between the two homologues is then

$$P(D) = \frac{L!}{D!(L - D)!} p^D(1 - p)^{L-D}. \tag{3.23}$$

The average value of D, $N(D)$ is thus pL, which becomes, upon substituting the values for p_i into (3.21) and (3.22),

$$N(D) = N_1'(x) + N_2'(x) - \frac{4}{3}\frac{N_1'(x)N_2'(x)}{L} \tag{3.24}$$

for Case 1,

$$N(D) = N_1'(x) + N_2'(x) - 2\frac{N_1'(x)N_2'(x)}{L} \tag{3.25}$$

for Case 2.

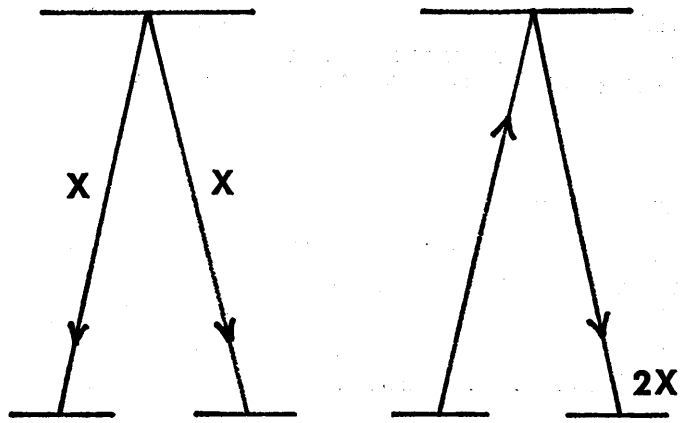
It should be noted that (3.24) and (3.25) are identical to (3.17) and (3.20), respectively, so that

$$N(D) = N_{2X}'(x). \tag{3.26}$$

Physically, this curious identity means it does not matter whether one considers each contemporary homologous DNA to have evolved from a common ancestral DNA over a time period such that each homologue has, on the average, received X hits, or whether one considers the homologues to have evolved one from the other during that same time period in such a way that the reference homologue has undergone $2X$ hits. The increased number of multiply hit sites and revertants in the latter case just equals the losses due to chance identity of base sites in the former. The two equivalent pathways are illustrated graphically in Figure 1.

3.7. *Approximations.* Under those conditions where the fraction to the right of $P(a_i)$ in (3.13) may be approximated by the Poisson distribution,

$$\frac{\sum_{x \leqq L} \sum_j N_{jx} n_{a_{ij}}(x)}{\sum_{a_i=1}^{X} \sum_{x \leqq L} \sum_j N_{jx} n_{a_{ij}}(x)} \approx P\left[\left(\frac{X}{L}\right), a_i\right] \equiv \frac{\exp\left\{-\frac{X}{L}\right\}\left[\frac{X}{L}\right]^{a_i}}{a_i!}. \tag{3.27}$$



$$2N'(x) - \frac{4}{3}\frac{N'(x)^2}{L} = N(D) = N'_{2X}(x)$$

FIGURE 1

Equivalent mutational pathways.

Equation (3.13) reduces, to the simple form

$$(3.28)\quad P = \sum_{a_i=1}^{\infty} \frac{1}{4}\left[1 + \frac{(-1)^{a_i}}{3^{a_i-1}}\right] \frac{\exp\left\{-\frac{X}{L}\right\}\left[\frac{X}{L}\right]^{a_i}}{a_i!} = \frac{1}{4}\left[1 + 3\exp\left\{-\frac{4}{3}\frac{X}{L}\right\}\right]$$

for Case 1. Equation (3.14) thus becomes

$$(3.29)\quad N'(x) = \frac{3}{4}\left[1 - \exp\left\{-\frac{4}{3}\frac{X}{L}\right\}\right]L$$

for Case 1, and

$$(3.30)\quad N'(x) = \frac{1}{2}\left[1 - \exp\left\{-2\frac{X}{L}\right\}\right]L$$

for Case 2. Equations (3.24) and (3.25) then become

$$(3.31)\quad N(D) = \frac{3}{4}\left[1 - \exp\left\{-\frac{8}{3}\frac{X}{L}\right\}\right]L$$

for Case 1, and

$$(3.32)\quad N(D) = \frac{1}{2}\left[1 - \exp\left\{-4\frac{X}{L}\right\}\right]L$$

for Case 2.

Equations (3.29), (3.30), (3.31), and (3.32), aside from being useful in their own right, form a convenient, rapid check on the numerical accuracy of more

exact calculations. The approximate equations are very handy computationally: a table of exponentials suffices to solve them.

In a somewhat different, but related, context, (3.29) has been independently derived by Neyman [19].

4. Derivations for proteins

4.1. *General.* The number of amino acid differences A between a present day protein and its ancestral homologue may be less than the average number of nucleotide base differences $N'(x)$ between the corresponding nucleic acids which code for these proteins for two reasons: first, amino acid codon degeneracy; and second, several of the differing nucleotide bases may fall within the same codon triplet.

4.2. *Amino acid codon degeneracy.* Some amino acids are coded for by more than one nucleotide triplet, so that a mutagenic event occurring within this triplet need not lead to an amino acid substitution. A recent tabulation of these codon triplets may be found in Watson [24]. If we consider a specific triplet coding for a particular amino acid, a *single* amino acid substitution will result at this position if either one, two, or three of the nucleotides of the triplet undergo mutation which results in the formation of a new triplet which codes for an amino acid differing from that in the ancestral protein. In Section A.1 of the Appendix the probability that a given amino acid will mutate to another amino acid if exactly one, two, or three nucleotide bases in the triplet coding for that amino acid are altered is calculated on the assumption that each of the sixty four codon triplets is equally probable. Although the three chain terminating triplets are necessarily less probable, that they sometimes do survive and find experimentally observable expression has been demonstrated by Weigert and Garen [25] (see also [11]). To the extent that these three codons cause deviations from randomness, this will show up in the viability parameter β (see equation (6.1)) or in the necessity to reduce X in order to obtain agreement with experiment. Experimental evidence that this assumption is approximately true has been provided by King and Jukes' analysis of known polypeptide sequences [13]. As in the preceding section, figures are quoted for both the case where any nucleotide base is free to mutate to any other (Case 1) and for the case where purines and pyrimidines may mutate only to a purine or pyrimidine (Case 2), respectively. These probabilities are when exactly one, two, or three nucleotides within a triplet have been altered, 0.7604, 0.9826, and 0.9931, respectively, for Case 1, and 0.6563, 0.9688, and 1.0000, respectively, for Case 2. Briefly, an amino acid substitution is almost certain when any combination of mutagenic events occurs within the amino acid codon, except for a single altered nucleotide in the third codon position. The above figures demonstrate that there is no large advantage with respect to conserving structure for an organism to develop a mechanism in which nucleotide base mutations are limited to purine to purine or pyrimidine to pyrimidine.

In order to complete the quantitative evaluation of the effect of codon degeneracy on the number of amino acid substitutions, we must calculate exactly how many codon triplets have sustained exactly one, exactly two, and exactly three altered (relative to the ancestral homologue) nucleotide bases. This problem is tackled in the next section.

4.3. *Multiple mutations within the same triplet.* If we consider a specific triplet coding for a particular amino acid position, a single amino acid substitution will result at this position if either one, two, or three of the nucleotides of the triplet undergo mutation, *provided* the number of single, double, and triple mutations are weighted by the appropriate probability for mutation from the preceding section. In particular, the average number of amino acid substitutions $N(A)$ is a function $F[N'(x), T]$, where T is the total number of amino acid residues in the homologue under consideration. At this point, it should be emphasized that it is *not* permissible to treat each nucleotide site as independent from the other sites. To do so will, in general, underestimate the number of amino acid substitutions. The reason for this is that $N'(x)$ already is corrected for multiple mutations at the same nucleotide site. The nature of the function F is complicated and best given by example in Section A.3 of the Appendix; its calculation is, however, straightforward. Some indication of its general properties may be had from the following considerations.

Represent the T sets of triplets from $L = 3T$ nucleotides in T rows as follows:

Row 1 _ _ _
Row 2 _ _ _
⋮
Row T _ _ _ _.

An amino acid substitution is represented by any row with one, two, or three altered nucleotides. For example, if $N'(x) = 7$, and $L = 18$,

x x x	x x _
x x _	x x _
x _ _	x x _
x _ _	x _ _
_ _ _	_ _ _
_ _ _	_ _ _
Form 1	Form 2

each represent four amino acid substitutions. We shall call such an array a *form*. A form is defined by stating the number of rows having zero, one, two, and three altered nucleotides, respectively, that is,

_ _ _	x _ _
_ x _	x _ x
x x x	_ x x
_ _ _	_ _ _
x _ _	_ _ _
_ x x	x x _

belong to Form 1 and Form 2, respectively. First we write down all possible forms. For a given $N'(x)$ and T, there can be A amino acid substitutions, where

$$I + \varepsilon \leqq A \leqq S, \tag{4.1}$$

$\varepsilon = \varepsilon(p)$: $\varepsilon(0) = 0$, $\varepsilon(1) = \varepsilon(2) = 1$, I and p are integers defined by the congruence

$$N'(x) \equiv p \pmod 3 \tag{4.2}$$

where $N'(x) = 3I + p$, and S is the smaller of $N'(x)$ and T. For example, if $N'(x) = 7$, $I = 2$, $p = 1$, $\varepsilon = 1$, and $S = 6$: $3 \leqq A \leqq 6$, that is, three, four, five, or six amino acid substitutions are possible. Before writing down any forms, A should be calculated since this will permit one to disregard those forms which are irrelevant to the problem, in this case any form with only one or two occupied rows. One must now count the number of ways W_i that the ith form can be realized. Let a_{0i}, a_{1i}, a_{2i} and a_{3i} be the number of rows in a given form which has zero, one, two and three altered nucleotides, respectively. Then, for that form,

$$\frac{N'(x)!T!3^{a_{1i}+a_{2i}}}{\prod_j^3 a_{ji}!}. \tag{4.3}$$

A convenient check on the calculations is provided by the fact that

$$\sum_{\text{all } i} W_i = \frac{L!}{[L - N'(x)]!}. \tag{4.4}$$

In the absence of codon degeneracy, W_i would be the number of ways of realizing, for a given form, exactly A_i amino acid substitutions, where

$$A_i = a_{1i} + a_{2i} + a_{3i}. \tag{4.5}$$

However, because of codon degeneracy, W_i must be reduced by approximately the factor f_i, where

$$f_i = 0.7604f_{1i} + 0.9826f_{2i} + 0.9931f_{3i} \tag{4.6}$$

for Case 1,

$$f_i = 0.6563f_{1i} + 0.9688f_{2i} + 1.0000f_{3i} \tag{4.7}$$

for Case 2, and

$$f_{ji} = \frac{a_{ji}}{A_i}. \tag{4.8}$$

The numerical coefficients for the f_{ji} were taken from Section A.1 of the Appendix. To find the number of ways W_A in which exactly A amino acid substitutions can occur, one adds together the number of ways of realizing *all* those forms which have *exactly* A rows occupied by *at least* one altered nucleotide:

$$W_A = \sum_{i=i_A} f_i W_i, \tag{4.9}$$

where the subscript A on i_A is to remind us that the summation is over *only* those forms having exactly A rows occupied. The probability that exactly A

amino acid substitutions have occurred between the ancestral and present day homologue is thus about

$$P(A) = \frac{W_A}{\sum_{\text{all } i} W_i}, \tag{4.10}$$

and the average number of amino acid substitutions is

$$N(A) = F[N'(x), T] = \sum_A AP(A). \tag{4.11}$$

Before leaving this section, we note that another quantity that is sometimes of interest is the proportion of amino acid substitutions that have occurred by one base, two base, and three base changes. This proportion p_j, $j = 1$, 2, or 3, is approximately,

$$p_j = \frac{\sum_i f_{ji} c_j W_i}{\sum_i f_i W_i}, \tag{4.12}$$

where the c_j are the coefficients of the f_{ji} in (4.6) and (4.7).

For the ith form, (4.6) and (4.7) have the effect of excluding those ways which represent fewer than A_i amino acid substitutions because of degenerate rows having one altered nucleotide *only* or two altered nucleotides *only* or three altered nucleotides *only*. Equations (4.6) and (4.7) do not exclude those ways which represent fewer than A_i amino acid substitutions because of combinations or degenerate rows of mixed type. For example, ways in which two rows of a form are degenerate, one of the rows containing one altered nucleotide, the second row containing either two or three altered nucleotides are not excluded.

For this reason W_A and $P(A)$ in equations (4.9) and (4.10) should be overestimates. However, since those ways not excluded by (4.6) and (4.7) represent amino acid substitutions fewer than A_i, those $P(A)$ and W_A with A less than A_i will be underestimated. If the total number of forms being considered is reasonably large, these two opposite effects will partially offset one another. Thus, despite the above limitations $N(A)$ in (4.11) is reasonably accurate because the average is taken over all forms. In equations (4.9), (4.10), and (4.12) W_A, $P(A)$, and p_j are less accurate because the summations are over fewer forms.

We are currently attempting to find a method of treating multiple hits within the same codon that is both less cumbersome and more exact than the method given in this section. Nevertheless, the above treatment is an improvement over existing methods which virtually ignore the problem.

4.4. *Calculation of $N(d)$: the average number of amino acid differences between two present day homologues.* The number of amino acid differences d between two present day homologues may be less than $N(A)$ because though each of the homologues may differ at a particular site from the ancestral homologue, the two homologues themselves may have the same amino acid at that site. In Section 3.6 of this paper, it was shown that the general form of the equation necessary to correct for this accidental coincidence between two present day sites is

$$N(d) = N_1(A) + N_2(A) - \alpha \frac{N_1(A)N_2(A)}{T}, \tag{4.13}$$

where the subscripts refer to the two present day homologues in question, and α is a numerical constant that depends only on the structure of the genetic code. This constant can be evaluated by considering two homologues that have evolved at the same rate for a sufficiently long time so that all sequences have been randomized. Then, $N(d) = N_1(A) = N_2(A)$, and

$$\alpha = \frac{1}{[N(A)/T]_{\text{equil}}} = \frac{1024}{963} = 1.0633, \tag{4.14}$$

because the probability, after randomization, that two homologous proteins will have Arg, Ser, or Leu at the same site is $3(6/64)^2$; that both homologues will have Ala, Thr, Gly, Val, or Pro is $5(4/64)^2$; that both will have Ile or Term is $2(3/64)^2$; that both will have Lys, His, Cys, Glu, Gln, Asp, Asn, Tyr, or Phe is $9(2/64)^2$; and that both will have Trp or Met is $2(1/64)^2$. Adding these probabilities together gives the probability that two present day homologues will both have the same amino acid at a given site. This probability is $244/64^2$. The probability that these two homologues will differ at a given site is

$$\left[1 - \frac{244}{64^2}\right] = \frac{963}{1024} = \left[\frac{N(A)}{T}\right]_{\text{equil}}. \tag{4.15}$$

Finally,

$$N(d) = N_1(A) + N_2(A) - \frac{1024}{963}\frac{N_1(A)N_2(A)}{T}. \tag{4.16}$$

The probability that exactly d amino acid substitutions will be observed between two present day homologous proteins is

$$P(d) = \frac{T!}{d!(T-d)!} p^d(1-p)^{T-d}. \tag{4.17}$$

4.5. *The "$N(d)/2$" approximation.* The number of amino acid substitutions between an ancestral protein and each of the two present day homologues is sometimes estimated by assuming $N_1(A) = N_2(A)$ and by assuming that their common value $N(A) = N(d)/2$. The first assumption may be experimentally checked by seeing if the number of amino acid substitutions between each homologue and a third evolutionary distant present day homologue are approximately equal. The exact relation between $N(A)$ and $N(d)$ may be found from (4.16):

$$\begin{aligned} N(A) &= \frac{T}{\alpha}\left[1 - \left(1 - \frac{\alpha N(d)}{T}\right)^{1/2}\right] \\ &= \frac{1}{2}N(d)\left[\begin{array}{l} 1 + \frac{1}{4}\frac{\alpha}{T}N(d) + \frac{1}{4}\cdot\frac{3}{6}\left(\frac{\alpha}{T}\right)^2 N(d)^2 \\ \quad + \frac{1}{4}\cdot\frac{3}{6}\cdot\frac{5}{8}\left(\frac{\alpha}{T}\right)^3 N(d)^3 \\ \quad + \cdots \end{array}\right]. \end{aligned} \tag{4.18}$$

Thus, if we are willing to accept an error in $N(A)$ no larger than ten per cent, the criterion for the validity of the "$N(d)/2$" approximation becomes

$$N(d) \leqq \frac{0.4T}{\alpha}. \tag{4.19}$$

5. Measures of error

5.1. *Measures of error in* (ND), $N'(x)$, $N(x)$, $N(A)$, *and* $N(d)$. The theoretically calculated value of $N(D)$ tends to be insensitive to small variations in X, the total number of mutagenic events, irrespective of their source (that is, whether they are statistical variations or nonstatistical ones). The reason for this is that an increase in X is partially offset by a compensating decrease due to a greater number of multiple hits, and increased back mutation. An analogous argument holds for small decreases in X. The same considerations hold for $N'(x)$ and $N(x)$. These considerations, of course, do not apply when $X \ll L$ or $X \gg L$.

The probability distributions that have been derived in this paper $P(x)$ (3.2) and (3.12), $P(a_i)$ (Table II), and $P(D)$ (3.23) are all well defined. Their frequency distributions therefore each possess a unique variance (second moment) that can, for a given X and L, be calculated in a straightforward manner by standard statistical methods (Hoel [8]), and this variance is a quantitative measure of the deviation from the average values $N(x)$, $N'(x)$, and $N(D)$ that one might expect to find in practice. This variance is particularly easy to calculate only for $P(D)$ and is given by

$$\sigma_D = \left(N(D)\left[1 - \frac{N(D)}{L}\right]\right)^{1/2}. \tag{5.1}$$

The true variance will be somewhat greater than this because the contribution from the variance of $N'(x)$ has been ignored in (5.1). Equation (5.1) shows that when $X \ll L$ or $X \gg L$ the error in $N(D)$ will be small. These are precisely the instances when we need an estimate of the error most badly, for in these instances the compensatory mechanisms that were discussed in the preceding paragraph do not apply. On the other hand, when X is of the order of L, these compensatory mechanisms do apply, so that (5.1) should still give a reasonably valid estimate of the error in $N(D)$, because of the insensitivity of the latter towards small variations.

Similarly, the standard deviation of the distribution of (4.17) is, closely,

$$\sigma_d = \left(N(d)\left[1 - \frac{N(d)}{T}\right]\right)^{1/2}. \tag{5.2}$$

The actual standard deviation of d will be somewhat greater than this because the variance of $N(A)$ has been ignored in (5.2).

6. Discussion

6.1. *General.* The only assumption made in deriving the statistics in this paper is that of *spatial* randomness along L. No assumptions about time are

involved, and the statistics remain valid for *any* particular time dependence of X, linear or nonlinear, random or nonrandom. No assumptions have been made about the number of mutagenic events that have occurred in each of several homologues. They may be the same or different. The statistics can thus handle the case of homologous macromolecules which have evolved at different rates.

In Sections A.2 and A.3 of the Appendix and Table I, it is demonstrated that to neglect the phenomena of multiple hits, back mutation, hits within the same codon, the degeneracy of the genetic code, and accidental (chance) coincidence between two homologous sites, can, in actual experimental cases, lead to errors in the value of $N(d)$ of a factor of 3.5. Obviously, with errors of this magnitude, it is impossible to construct with confidence any meaningful phylogenetic trees, or to conclude that a series of sequence homologies did or did not arise by a stochastic pathway. The formulas in this paper permit one to quantitatively correct for the above phenomena so that unambiguous answers to the questions can be given.

If the observed number of differences between two homologous macromolecules differ from the calculated value of $N(D)$ or $N(d)$ by much more than twice the statistical error ((5.1) and (5.2)), the cause of such discrepancy may lie with one of the following factors.

(1) The spatial distribution of the mutagenic events is nonrandom along L. In fact, if the other factors below can be eliminated as causes of discrepancy, the statistics of this paper may be used as an algorithm to search for nonrandomness *within* a single molecule by considering subsegments of that molecule which contain $\ell < L$ nucleotide bases.

(2) $N(D)$ or $N(d)$ are not the appropriate statistic. These are an average value of D and d. Other values can and will occur with a relative frequency given by (3.23) and (4.17). If for some applications, an investigator wishes to utilize the most probable, rather than the average, values of x, x', D and d, these most probable values may be calculated from the equations given in this paper.

(3) The input data is incorrect, that is, one's estimate of X is wrong; this corresponds to incorrect assumptions, for example, linearity or randomness in time, about the mutation rate. If after all other causes have been eliminated, a discrepancy still remains, one should seriously consider revising the numerical value of the mutation rate, as well as assumptions about its temporal dependence.

(4) Viability—this is the least well known of all the quantities and may be one cause of any nonrandomness falling under category (1) above. Consider several homologous macromolecules and in particular that region of each for which one has calculated $N'(x)$. Call the number of sites in this region which for *all* the macromolecules contain the same nucleotide base at a given site. (Different sites may contain different nucleotide bases.) The viability β is defined as

$$\beta = 1 - \frac{f}{L}. \tag{6.1}$$

This may or may not be a good estimate of the viability depending on the number of homologues available. All calculations are then repeated starting with a

revised estimate of X and L, namely, βX and βL to see if improved agreement results. This procedure should be followed only *after* the first three factors above have been adequately accounted for: otherwise β becomes no more than a "fudge factor."

6.2. *Minimum mutation distance.* The relationship between the *minimum mutation distance* [4] and the calculations in the present paper can be made clear by considering the proportion of amino acid substitutions that have occurred by actual one base, two base, and three base changes: p_1, p_2, and p_3 (Appendix, Section A.3). In the sense implied by the concept of minimum mutation distance, a three base change can occur in two ways: (a) for one present day homologue each base of the codon triplet differs from the corresponding base of the ancestral homologue, or (b) for one present day homologue two bases of the codon triplet differ from the corresponding ancestral bases, and for the second present day homologue the corresponding two bases are identical to those in the ancestral homologue while the remaining third base differs from the ancestral homologue. Thus, the proportion of amino acid substitutions that have a minimum mutation distance of 3 ($P_{\text{M.M.D.}=3}$) is very nearly,

$$P_{\text{M.M.D.}=3} = 0.05(2)\left[p_3\left(1 - \frac{N(A)}{T}\right) + \frac{1}{3}p_1p_2\frac{N(A)}{T}\right]. \tag{6.2}$$

The factor 0.05 arises from the fact that 95 per cent of actual three base changes are "silent" because the algorithm by which minimum mutation distance is computed counts these "silent" 3 base changes as either 1 or 2 base changes. If the numerical values of p_1, p_2, p_3, $N(A)$, and T, which are given in A.3 of the Appendix are substituted into (6.2), then $P_{\text{M.M.D.}=3} = 0.0078$. Equation (6.2) is an underestimate for it neglects all cases where the sum of the total number of base changes for a given codon is greater than three (the sum is taken over both homologues). Analogous calculations may be made to find $P_{\text{M.M.D.}=2}$ and $P_{\text{M.M.D.}=1}$.

Alternatively, one can utilize the principles embodied in (3.15) through (3.17) and calculate

$$P_{X,\text{M.M.D.}=3} = 0.05p_{2X,3}. \tag{6.3}$$

In either case, it is clear that, except for very low mutation rates, the minimum mutation distance bears a minimum relationship to the true course of events.

6.3. *Application to experimental data.* In Table IV are shown the homologous sequence fragments of the A fibrinopeptides of the sheep, goat, ox, and reindeer, with which the numerical calculations will be compared. These short fragments were chosen for two reasons: first, they contain a constant region of three residues, and it is of interest whether the theoretical methods can detect this region; and second, for the variable region the minimum mutation rate estimated by Doolittle and Blombaeck [2] as 10^{-7} mutagenic events/year/codon may be as rapid as the rate of evolution of DNA itself (Kohne [14]), and thus the deficiencies of the "minimum mutation distance" show up most clearly. The fragments were kept short so as not to initially become bogged down in calculational irrele-

TABLE IV

ILLUSTRATIVE FIBRINOPEPTIDE A FRAGMENTS FOR COMPARISON WITH CALCULATED VALUES

See Section 6.3 for explanation.

Organism	Sequence	Amino acid differences	M.M.D.
Goat:	(Asp-Ser-Asp-Pro-Val-Gly)		
		0	0
Sheep:	(Asp-Ser-Asp-Pro-Val-Gly)		
		3	4
Ox:	(Gly-Ser-Asp-Pro-Pro-Ser)		
		2	2
Reindeer:	(Gly-Ser-Asp-Pro-Ala-Gly)		
Amino Acid Position Number:	17 16 15 14 13 12		

vancies. For computational purposes (See Appendix, Section A.2) it is assumed that these four artiodactyls diverged from their most recent common ancestor 15 million years ago giving a total of 9 ($10^{-7} \times 6 \times 15 \cdot 10^6$) primary mutagenic events randomly distributed over a region of 18 nucleotide bases or 6 amino acids. Thus, the ratio, hits:sites::9:18 is ½, a value chosen to avoid the trivial cases where the sites are saturated with hits or hardly hit at all. The actual time of divergence of the most distantly related pair (sheep-reindeer) may be closer to 30 million years [2]. If the latter figure is used the general conclusions are only strengthened.

The first row of Table I demonstrates both that the expected number, 5.14, of amino acid differences between the homologous fibrinopeptide fragments being compared is less than the number of primary mutagenic events by a factor of 3.5 and that the experimentally observed number of differences (0–3) is inconsistent with a stochastic mechanism. The second row of Table I shows that when the constant region of three residues is taken into account, agreement with experiment is obtained. The third and fourth rows of Table I demonstrate that the observed differences can also be explained by assuming a mutation rate 2/9 of that in the first two rows, or, 2.2×10^{-8} mutagenic events/year/codon. The values in parentheses in Table I illustrate the fact that at the level of nucleotide base or amino acid differences, it is not statistically possible, in fragments as short as those being considered here, to distinguish between unrestricted mutation, where any base may mutate to any one of the other three bases, and restricted mutation, where purine ↔ pyrimidine mutations are forbidden.

◇ ◇ ◇ ◇ ◇

The recursion formula for $m(X, x)$ (3.10) was developed in collaboration with Mr. Andrew Lebor during several evenings of discussion in the Spring of 1966. I wish to thank Dr. Thomas H. Jukes for pointing out that most three base

changes are "silent" when the experimental data are analyzed with the concept of minimum mutation distance and for providing a quantitative measure of this "silence."

◇ ◇ ◇ ◇ ◇

ADDENDUM

Since this paper was presented I have become aware, through the courtesy of Patricia Altham, University of Cambridge, Department of Pure Mathematics and Mathematical Statistics, that Equations (3.2) and (3.12) and (3.1) and (3.11) can be explicitly formulated as

$$P(x) = \frac{1}{L^X}\frac{L!}{x!(L-x)!}\sum_{\nu=0}^{x}(-1)^{\nu}\frac{x!}{\nu!(x-\nu)!}(x-\nu)^X \tag{7.1}$$

and

$$N(x) = L - L\left(1 - \frac{1}{L}\right)^X \tag{7.2}$$

with variance

$$\sigma^2(x) = L\left(\frac{L-1}{L}\right)^X\left[1 - \left(\frac{L-1}{L}\right)^X\right] + L(L-1)\left[\left(\frac{L-2}{L}\right)^X - \left(\frac{L-1}{L}\right)^{2X}\right]. \tag{7.3}$$

Similarly (3.14) can be explicitly written for Case 1

$$N'(x) = \frac{3}{4}L\left[1 - \left(1 - \frac{4}{3L}\right)^X\right] \tag{7.4}$$

with variance

$$\sigma^2(x) = \frac{3}{16}L\left[1 + 2\left(1 - \frac{4}{3L}\right)^X - 3\left(1 - \frac{4}{3L}\right)^{2X}\right] + \frac{9}{16}L(L-1)\left[\left(1 - \frac{8}{3L}\right)^X - \left(1 - \frac{4}{3L}\right)^{2X}\right], \tag{7.5}$$

and for Case 2

$$N'(x) = \frac{1}{2}L\left[1 - \left(1 - \frac{2}{L}\right)^X\right], \tag{7.6}$$

with variance

$$\sigma^2(x) = \frac{1}{4}L\left\{1 - \left(1 - \frac{2}{L}\right)^{2X} + (L-1)\left[\left(1 - \frac{4}{L}\right)^X - \left(1 - \frac{2}{L}\right)^{2X}\right]\right\}. \tag{7.7}$$

In a like manner, the exact variance, $\sigma^2(D)$, of D is given by (7.5) and (7.7) if X is replaced by $2X$ on the right side of these equations. The exact variance of D is somewhat less or greater than that given by the square of (5.1) depending on whether $L > 3$ or $L < 3$, respectively. The approximate nature of (5.1) results from the fact that (3.23) is an oversimplification of the true distribution of D because whether or not a difference occurs at a site is not really independent

of whether differences exist at other sites so long as the total number of mutagenic events $2X$ is fixed. Analogous considerations apply to (4.17) and (5.2).

The above formulas are computationally more convenient than the ones in the text. In addition, the quantitative expressions for the variances (7.3), (7.5), and (7.7) supplement the discussion in Section 5.

A good reference for the mathematical techniques that are needed to solve the "occupancy problems" that arise in studies of molecular evolution is William Feller's book [3].

Note added in proof. The "derivations for proteins" in Section 4 and the discussion on minimum mutation distance in Section 6 have since been made quantitatively exact rather than approximate by Holmquist, Cantor, and Jukes [30]. The methods described here have also been applied to analyze evolutionary changes in the cytochrome *c* globins and immunoglobulins by Jukes and Holmquist [31].

APPENDIX

A.1. Probabilities for amino acid mutation

Probabilities for amino acid mutation are given in Tables AI, AII, and AIII. In Table AI, the calculation for the probability corresponding to Ser for Case 1 is

$$\tfrac{25}{27} = \tfrac{2}{3}[1 - \tfrac{1}{3}\cdot\tfrac{1}{3}\cdot\tfrac{1}{2}] + \tfrac{1}{3}[1 - \tfrac{1}{3}\cdot\tfrac{1}{3}(1)]. \tag{A.1.1}$$

The computation for the averaged probability is

$$0.9931 = \tfrac{1}{64}\,[(6)\tfrac{25}{27} + 58(1)]. \tag{A.1.2}$$

These examples indicate the general method of computation. In the remaining tables only the results are given.

A.2. Illustrative example: DNA

What follows is a detailed numerical analysis of the A fibrinopeptides of the ox, sheep, goat, and reindeer. These peptides contain 19 amino acids. Doolittle

TABLE AI

PROBABILITIES FOR AMINO ACID MUTATION
ALL 3 CODON POSITIONS ALTERED
$\underline{x}\ \underline{x}\ \underline{x}$

Amino acid	Probability Case 1	Case 2
Ser	25/27	1
All others	1	1
Average probability	0.9931	1

TABLE AII

PROBABILITIES FOR AMINO ACID MUTATION
ANY 2 CODON POSITIONS ALTERED
x x _, _ x x, x _ x

Term indicates chain terminating codon: UAA, UAG, or UGA.

Type of change	Amino acid	Probability Case 1	Case 2
x x _	Ser	25/27	1
	All others	1	1
	Average probability	0.9931	1
_ x x	Term	25/27	1/3
	All others	1	1
	Average probability	0.9965	0.9688
x _ x	Arg	7/9	1
	Leu	7/9	1/3
	All others	1	1
	Average probability	0.9583	0.9375
	Averaged average probability	0.9826	0.9688

TABLE AIII

PROBABILITIES FOR AMINO ACID MUTATION
ANY 1 CODON POSITION ALTERED
x _ _, _ x _, _ _ x

Type of change	Amino acid	Probability Case 1	Case 2
x _ _	Arg	7/9	1
	Leu	7/9	1/3
	All others	1	1
	Average probability	0.9583	0.9375
_ x	Term	7/9	1/3
	All others	1	1
	Average probability	0.9896	0.9688
_ _ x	Met	1	1
	Trp	1	1
	Term	7/9	1/3
	Lys	2/3	0
	His	2/3	0
	Asp	2/3	0
	Asn	2/3	0
	Glu	2/3	0
	Gln	2/3	0
	Cys	2/3	0
	Tyr	2/3	0
	Phe	2/3	0
	Ile	1/3	1/3
	Arg	2/9	0
	Ser	2/9	0
	Leu	2/9	0
	All others	0	0
	Average probability	0.3333	0.0625
	Averaged average probability	0.7604	0.6563

and Blombaeck [2] have estimated a minimum mutation rate for these peptides of 10^{-7} mutations/year/amino acid. For simplicity, we shall also assume that 10^{-7} mutations/year is the rate at which mutagenic events occur per nucleotide base triplet; the actual rate, of course, will be slightly greater than this because of codon degeneracy. In particular, consider only that segment of six amino acids numbered 12 through 17 by the above authors. What is the number of differences to be expected between any two of the corresponding homologous DNA's which code for positions 12 through 17 after each homologue has had 15 million years to develop from a common ancestral DNA?

For the solution, first we have $L = 3 \times 6 = 18$ and $X = 10^{-7} \times 15 \cdot 10^{6} \times 6 = 9$. In general X will be nonintegral. In such a case one carries through the calculations for the integers on either side of X and at the end takes a weighted average.

Second, we need to list all $x \leqq 18$ part partitions of 9 as shown in Table AIV. We use the *Table of Partitions* [7] in order to find the number of partitions $P(9, x)$. The physical meaning of the particular partition 5, 4 is that two nucleotide bases and only two have been altered. One has been hit a total of five times, the other only four. The computation of the column N_{jx} is illustrated in the case when $j = 5, 2, 2$ and $x = 3$:

$$N_{jx} = \frac{9!}{(1!2!)(5!2!2!)} = \frac{18!}{15!} = 1{,}850{,}688. \tag{A.2.1}$$

We note that

$$p(9, 9) = \sum_{x=1}^{18} P(X, x) = 30, \tag{A.2.2}$$

so that we have left no partitions out, and that

$$L^X = 18^9 = \sum_{x \leqq 18} \sum_{j} N_{jx} = 1{,}980 \times 10^8, \tag{A.2.3}$$

so that our summation of the N_{jx} is correct. Thus, $P(X, x)$ and $p(X, X)$, from [7], and L^X provide independent checks on the accuracy of the calculations. The probability that exactly one, two, $\cdots$, nine sites have been hit is thus

$$\begin{aligned}
&P(1) = 18/1{,}980 \times 10^8 = 0.000,\\
&P(2) = (3{,}054 + 11{,}016, +25{,}704 + 38{,}556)/1{,}980 \times 10^8 = 0.000,\\
&P(3) = 0.000,\\
&P(4) = 0.002,\\
&P(5) = 0.036,\\
&P(6) = 0.178,\\
&P(7) = 0.373,\\
&P(8) = 0.321,\\
&P(9) = 0.089.
\end{aligned} \tag{A.2.4}$$

The *most probable* value of x is therefore 7 and the average value is

$$\begin{aligned}
N(x) &= 9(0.089) + 8(0.321) + 7(0.373)\\
&\qquad + 6(0.178) + 5(0.036) + 4(0.002)\\
&= 7.24.
\end{aligned} \tag{A.2.5}$$

TABLE AIV

PARTITIONS OF 9 FOR $x \leqq 18$ AND COMPUTATIONS FOR N_{jx}, THE NUMBER OF WAYS OF REALIZING $(x, X)_j$

	Partition	$P(9, x)$	N_{jx}
(1, 9):	9	1	18
(2, 9):	8, 1		3,054
	7, 2		11,016
	6, 3		25,704
	5, 4	4	38,556
(3, 9)	7, 1, 1		176,256
	6, 2, 1		1,233,792
	5, 3, 1		2,467,548
	5, 2, 2		1,850,688
	4, 4, 1		1,542,240
	4, 3, 2		6,168,960
	3, 3, 3	7	1,370,880
(4, 9)	6, 1, 1, 1		0×10^8
	5, 2, 1, 1		1
	4, 3, 1, 1		1
	4, 2, 2, 1		1
	3, 3, 2, 1		0
	3, 2, 2, 2	6	1
(5, 9)	5, 1, 1, 1, 1		1
	4, 2, 1, 1, 1		13
	3, 3, 1, 1, 1		9
	3, 2, 2, 1, 1		39
	2, 2, 2, 2, 1	5	10
(6, 9)	4, 1, 1, 1, 1, 1		17
	3, 2, 1, 1, 1, 1		168
	2, 2, 2, 1, 1, 1	3	168
(7, 9)	3, 1, 1, 1, 1, 1, 1		134
	2, 2, 1, 1, 1, 1, 1	2	605
(8, 9)	2, 1, 1, 1, 1, 1, 1, 1	1	635
(9, 9)	1, 1, 1, 1, 1, 1, 1, 1, 1	1	176

From this example, it is clear that one need only write out those partitions whose probabilities are high. In the present case, it is sufficient to calculate only the 5 through 9 part partitions of 9. This reduces the number of N_{jx} which must be calculated from 30 to 12, a considerable saving in time and effort. Those partitions which *do* have high probability may be found *prior* to writing out any of the partitions by using (3.10) and (3.12). We have

$$\begin{aligned} m(9, 1) &= 1 \\ m(9, 2) &= 2^9 - \frac{2!}{1!1!}(1) = 510 \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots . \end{aligned} \tag{A.2.6}$$

If these values of $m(X, x)$ are substituted into (3.12), we get

$$\begin{aligned} P(1) &= 1(18!)/18^9(1!)(17!) = 0.000 \\ P(2) &= 510(18!)/18^9(2!)(16!) = 0.000 \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \end{aligned} \tag{A.2.7}$$

These values agree exactly with those calculated by the longer method.

TABLE AV

COMPUTATION OF THE PROBABILITY OF EXACTLY a_i HITS AND OF THE PROBABILITY $P(a_i)$ THAT THE FINAL NUCLEOTIDE IS THE SAME AS THE ORIGINAL NUCLEOTIDE

a_i	$\sum_{x \leqq L} \sum_j N_{ij} n_{a_{ij}}(x)$	Probability of exactly a_i hits	$P(a_i)$ Case 1	$P(a_i)$ Case 2
1	11.11×10^{11}	.784	.000	0
2	2.66	.188	.333	1
3	.36	.025	.222	0
4	.03	.002	.259	1
5	.00	.000	.247	0
6	.00	.000	.250	1
7	.00	.000	.250	0
8	.00	.000	.250	1
9	.00	.000	.250	0
	14.16×10^{11}			

Third, to find the effect of back mutation, we construct Table AV. As a sample calculation for column 2 let $a_i = 1$. Then we have

$$\begin{aligned} N_{1j} n_{1j}(x) = {} & 0(18) + 1(3{,}054) + [2(176{,}256) \\ & + 1(1{,}233{,}792) + 1(2{,}467{,}548) + 1(1{,}542{,}240)] \\ & + 10^8[3(0) + 2(1) + 2(1) + 1(1) + 1(0)] \\ & + 10^8[4(1) + 3(13) + 3(9) + 2(39) + 1(10)] \\ & + 10^8[5(17) + 4(168) + 3(168)] \\ & + 10^8[6(134) + 5(605)] \\ & + 10^8 \cdot 7(635) + 10^8 \cdot 8(176) \\ = {} & 11.11 \times 10^{11}. \end{aligned} \tag{A.2.8}$$

To calculate the probability that a site has been hit exactly a_i times for $a_i = 1$, we have $0.784 = 11.11 \times 10^{11}/(14.16 \times 10^{11})$.

The *most probable* probability for back mutation at a site is therefore,

$$P(2) = 0.188 \times 0.333 = 0.0626 \tag{A.2.9}$$

for Case 1, and

$$P(2) = 0.188 \times 1 = 0.188 \tag{A.2.10}$$

for Case 2, and the average probability for back mutation at a site is

$$(A.2.11)\quad P = 0(0.784) + 0.188(0.333) + 0.025(0.222) + 0.002(0.259) = 0.0687$$

for Case 1, and

$$(A.2.12)\quad P = 0(0.784) + 0.188(1) + 0.025(0) + 0.002(1) = 0.1884.$$

Fourth, since in the present instance there is no large difference between the most probable values and the average values, we shall continue the calculations with the average values only. The average number of altered nucleotides in each homologue will be,

$$(A.2.13)\quad N'(x) = (1 - 0.0687)7.24 = 6.74$$

for Case 1, and

$$(A.2.14)\quad N'(x) = (1 - 0.1884)7.24 = 5.88$$

for Case 2.

Fifth, now we can calculate the average number of nucleotide base differences between the two homologues. Assuming for simplicity that each has evolved from the ancestral polynucleotide at roughly equal rates, we have

$$(A.2.15)\quad N(D) = 6.74 + 6.74 - \frac{4}{3}\frac{(6.74)(6.74)}{18} = 10.12 \pm 1.80(\sigma)$$

for Case 1, and

$$(A.2.16)\quad N(D) = 5.88 + 5.88 - 2\frac{(5.88)(5.88)}{18} = 7.91 \pm 2.04(\sigma)$$

for Case 2.

This completes the numerical calculations. Had no corrections of any sort been made, the incorrectly calculated value of $N(D)$ would have been $9 + 9 = 18$, an error of the order of 103 per cent. Had corrections been made for multiple hits at the same nucleotide site alone, or for chance coincidence between homologous sites alone, the errors would have been of the order of 63 per cent and 16 per cent, respectively. Taking into account multiple hits and chance coincidence, but not back mutation, reduces the error to about 7 per cent. Attention should also be directed to the magnitude of the statistical error. The number of observed differences between the two homologues in this particular example could lie anywhere between 5 to 13 and still justify the statistical conclusion that the two homologues are each derived from a common ancestral polynucleotide. In view of this wide range within which it is not possible to reasonably conclude that the hypothesis of common ancestry is false, it is all the more important that all possible corrections be made before conclusions are drawn about the phylogeny of two homologues.

A.3. Illustrative example: proteins

In Section A.2 the number of expected differences between any single present day homologue and the ancestral DNA coding for these positions was found to be $N'(x) = 6.74$ and 5.88 for unrestricted (Case 1) and restricted (Case 2) mutation, respectively. Because the DNA sequences are not experimentally available for these peptides, we here continue the calculations by using the methods described in the main body of this paper (Section 4) to find $N(d)$, the number of amino acid differences expected between two present day homologues. In what follows, to conserve space, we shall show the calculations for Case 1 and quote them for Case 2.

Because $N'(x)$ is not integral (6.74), we must calculate $N(A)$ for both $N'(x) = 6$ and $N'(x) = 7$. Detailed calculations are given only for $N'(x) = 7$

TABLE AVI

CALCULATIONS FOR $N(A)$

Case: $N'(x) = 7$; $I = 2$, $p = 1$; $\varepsilon = 1$; $3 \leqq A \leqq 6$; $T = 6$.
Details of computations: $W_i = 907{,}200 = (7!6!3^{1+0})/(3!1!0!2!)$;
$f_i = 0.9155 = (\frac{1}{3})(0.9155) + (0)(0.9826) + (\frac{2}{3})(0.9931)$;
$f_iW_i = F_iW_i = 830{,}542 = 0.9155(907{,}200)$.

	$A = 3$		$A = 4$		$A = 5$		$A = 6$
	$\underline{x}\ \underline{x}\ \underline{x}$	$\underline{x}\ \underline{x}\ \underline{x}$	$\underline{x}\ \underline{x}\ \underline{x}$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ \underline{x}\ \underline{x}$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ \underline{x}\ -$
	$\underline{x}\ \underline{x}\ \underline{x}$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ -\ -$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ -\ -$
	$\underline{x}\ -\ -$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ -\ -$	$\underline{x}\ \underline{x}\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$
	$-\ -\ -$	$-\ -\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$
	$-\ -\ -$	$-\ -\ -$	$-\ -\ -$	$-\ -\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$	$\underline{x}\ -\ -$
	$-\ -\ -$	$-\ -\ -$	$-\ -\ -$	$-\ -\ -$	$-\ -\ -$	$-\ -\ -$	$\underline{x}\ -\ -$
W_i	907,200	2,721,600	24,494,400	24,494,400	12,247,200	73,483,200	22,044,960
f_{1i}	$\frac{1}{3}$	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{4}{5}$	$\frac{3}{5}$	$\frac{5}{6}$
f_{2i}	0	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{3}{4}$	0	$\frac{2}{5}$	$\frac{1}{6}$
f_{3i}	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	0	$\frac{1}{5}$	0	0
f_i	0.9155	0.9861	0.8741	0.9270	0.8069	0.8492	0.7974
f_iW_i	830,542	2,683,770	21,410,555	22,706,309	9,882,266	61,938,989	17,578,651

$W_{A=3} = 830{,}542 + 2{,}683{,}770 = 3{,}514{,}312$
$W_{A=4} = 44{,}116{,}864$
$W_{A=5} = 71{,}821{,}255$
$W_{A=6} = 17{,}578{,}651$
$W_A = 137{,}031{,}082$
$W_i = 160{,}392{,}960$

in Table AVI. The probabilities that exactly A amino acid substitutions have occurred between the ancestral and present day homologue are, from equation (4.10),

(A.3.1)
$$P(3) = \frac{3{,}514{,}312}{160{,}392{,}960} = 0.0219,$$
$$P(4) = 0.2750,$$
$$P(5) = 0.4477,$$
$$P(6) = 0.1096.$$

The average number of amino acid substitutions is thus,

(A.3.2) $N(A) = 3(0.0219) + 4(0.2750) + 5(0.4477) + 6(0.1096) = 4.06.$

The proportions of these substitutions that have occurred by one base, two base, and three base changes can be obtained using (4.12) and are

(A.3.3)
$$p_1 = 0.503,$$
$$p_2 = 0.424,$$
$$p_3 = 0.073,$$

where the calculation for p_1 is

(A.3.4)
$$p_1 = \frac{0.7604}{137{,}031{,}082}[\tfrac{1}{3}(907{,}200) + 0(2{,}721{,}600) + \tfrac{1}{2}(24{,}494{,}400) + \tfrac{1}{4}(24{,}494{,}400) + \tfrac{4}{5}(12{,}247{,}200) + \tfrac{3}{5}(73{,}483{,}200) + \tfrac{5}{6}(22{,}044{,}960)] = 0.503.$$

When similar calculations are made for $N'(x) = 6$, we find

(A.3.5)
$$N(A) = 3.66,$$
$$p_1 = 0.573,$$
$$p_2 = 0.377,$$
$$p_3 = 0.050.$$

Since $N'(x) = 6.74$ is 74/100 of the way between 6 and 7, the number of amino acid substitutions is

(A.3.6) $N(A) = 3.66 + 0.74(4.06 - 3.66) = 3.96,$

and

(A.3.7)
$$p_1 = 0.525,$$
$$p_2 = 0.412,$$
$$p_3 = 0.066.$$

A corresponding result for Case 2, where $N'(x) = 5.88$ is $N(A) = 3.28$. The average number of amino acid substitutions between two present day homologues is, from equation (4.16),

(A.3.8)
$$N(d) = 3.96 + 3.96 - \tfrac{1}{6}(3.96)(3.96)(1.0633) = 5.14 \pm 0.86$$

for Case 1, and

(A.3.9) $N(d) = 4.65 \pm 1.02$

for Case 2. This completes the calculations.

The observed number of amino acid substitutions between any pair of the fibrinopeptides A under discussion varies from zero (sheep-goat) to three (ox-

sheep, ox-goat); the pairs ox-reindeer, sheep-reindeer, and goat-reindeer each have two substitutions. The calculated value of 5.14 appears to be considerably too high, and its error ($\sigma = 0.86$) is too small to allow the difference to be explained as statistical fluctuation. One might be tempted to conclude that the correct value of the mutation rate is nearer 0.39×10^{-7} mutagenic events/year/codon rather than 1.0×10^{-7}, and indeed, this is a valid possibility (see Table I rows three and four; also, Section 6.3). However, the fact that the figure 1.0×10^{-7} represents a minimum estimate argues against this interpretation. A second reasonable interpretation emerges if we consider the possibility that the individual mutagenic events are spatially nonrandom along the nucleic acid segment coding for these peptides. Examining the actual amino acid sequence in positions 12 through 17, we find that positions 14 through 16 contain the same sequence in all these mammals, namely, H-Ser-Asp-Pro-Oh. Another possibility is that organisms which sustained mutations in these positions did not survive. In either case we can estimate the *viability* β from (6.1), $\beta = 1 - \frac{3}{6} = 0.5$. We now repeat the calculations of Sections 3 and 4 using the revised values of X and L, namely,

$$\begin{aligned} X' &= \beta X = 0.5(9) = 4.5, \\ L' &= \beta L = 0.5(18) = 9. \end{aligned} \tag{A.3.10}$$

When this is done we find that

$$N(d) = 2.62 \pm 1.51 \tag{A.3.11}$$

for Case 1, and

$$N(d) = 2.40 \pm 1.38 \tag{A.3.12}$$

for Case 2, in agreement with the experimental values found above.

REFERENCES

[1] M. O. Dayhoff, *Atlas of Protein Sequence and Structure*, Vol. 4, Silver Spring, Md., National Biomedical Research Foundation, 1969.

[2] R. F. Doolittle and R. Blombaeck, "Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications," *Nature*, Vol. 202 (1964), pp. 147–152.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, New York, Wiley, 1968, (3rd ed.).

[4] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees: A method based on mutation distances as estimated from cytochrome c sequences of general applicability," *Science*, Vol. 155 (1967), pp. 279–284.

[5] ———, "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution," *Biochem. Genet.*, Vol. 4 (1970), p. 5797.

[6] L. Gatlin, "The information content of DNA II," *J. Theor. Biol.*, Vol. 18 (1968), pp. 181–194.

[7] H. Gupta, C. E. Gwyther, and J. C. P. Miller, *Tables of Partitions, Royal Society Mathematical Tables*, Vol. 4, Cambridge, University Press, 1958.

[8] P. G. HOEL, *Introduction to Mathematical Statistics*, New York, Wiley, 1954, (2nd ed.). (See pp. 61, 74, 101.)

[9] W. R. HOLMQUIST, "The origin, partial structure and properties of Hemoglobins A_{Ic}," Ph.D. Thesis, California Institute of Technology, Pasadena, 1966. (See pp. 249–258.)

[10] T. H. JUKES, "Some recent advances in studies of the transcription of the genetic message," *Adv. Biol. Med. Phys.*, Vol. 9 (1963), pp. 1–41.

[11] ———, *Molecules and Evolution*, New York, Columbia University Press, 1966. (See Chapter 4.)

[12] M. KIMURA and T. OHTA, *On the rate of molecular evolution*, Mishima, Japan, National Institute of Genetics, 1970.

[13] J. L. KING and T. H. JUKES, "Non-Darwinian evolution," *Science*, Vol. 164 (1969), pp. 788–798.

[14] D. E. KOHNE, "Evolution of higher organism DNA," *Quart. Rev. Biophys.*, Vol. 33 (1970), pp. 327–375.

[15] E. MARGOLIASH, "The amino acid sequence of cytochrome *c* in relation to its function and evolution," *Can. J. Biochem.*, Vol. 42, No. 5 (1964), pp. 745–753.

[16] M. MARTIN and B. H. HOYER, "Adenine plus thymine and guanine plus cytosine enriched fractions of animal DNA's as indicators of polynucleotide homologies," *J. Molec. Biol.*, Vol. 27 (1967), pp. 113–129.

[17] H. MATSUBARA, T. H. JUKES, and C. R. CANTOR, "Structural and evolutionary relationship of ferredoxins," *Brookhaven Symp.*, Vol. 21 (1968), pp. 201–216.

[18] S. B. NEEDLEMAN and C. D. WUNSCH, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, Vol. 48 (1970), pp. 443–453.

[19] J. NEYMAN, "Molecular studies of evolution: A source of novel statistical problems," *Proceedings of the Purdue Symposium on Statistical Decision Theory*, Lafayette, Indiana, Purdue University Press, 1970.

[20] L. PAULING and E. ZUCKERKANDL, "Chemical paleogenetics, molecular "restoration studies" of extinct forms of life," *Acta Chem. Scan.*, Vol. 17, Suppl. #1 (1963), pp. S9–S16.

[21] T. A. REICHERT and A. K. C. WONG, "An application of information theory to genetic mutations & the matching of polypeptide sequences," *Biotechnology Program*, Pittsburgh, Carnegie-Mellon University, 1970, Personal communication.

[22] J. B. REID and G. MONTPETIT, *Tables of Factorials 0:–9999!*, Washington, D.C., National Academy of Sciences, National Council Publications 1039, 1962.

[23] C. E. SHANNON and W. WEAVER, *The mathematical theory of communication*, Urbana, The University of Illinois Press, 1949.

[24] J. D. WATSON, *Molecular Biology of the Gene*, New York, W. A. Benjamin, 1970, (2nd ed.). (See Chapter 13.)

[25] M. G. WEIGERT and A. GAREN, "Base composition of nonsense codons in *E. coli;* Evidence from amino-acid substitutions at a tryptophan site in Alkaline phosphatase," *Nature*, Vol. 206 (1965), pp. 992–994.

[26] A. C. WILSON and V. M. SARICH, "A molecular time scale for human evolution," *Proc. Nat. Acad. Sci. USA*, Vol. 63 (1969), pp. 1088–1093.

[27] E. ZUCKERKANDL, "The evolution of hemoglobin," *Sci. Amer.*, Vol. 189, May (1965), pp. 110–118.

[28] E. ZUCKERKANDL and L. PAULING, "Molecular disease, evolution, and genic heterogeneity," *Horizons in Biochemistry* (edited by M. Kasha and B. Pullman), New York, Academic Press, 1962, pp. 189–225.

[29] ———, "Evolutionary divergence and convergence in proteins," *Evolving Genes and Proteins* (edited by V. Bryson and H. J. Vogel), New York, Academic Press, 1965, pp. 97–166.

Added in proof.

[30] R. HOLMQUIST, C. CANTOR, and T. H. JUKES, "Improved procedures for comparing homologous sequences in molecules of proteins and nucleic acids," *J. Molec. Biol.*, Vol. 64 (1972), pp. 145–161.

[31] T. H. JUKES and R. HOLMQUIST, "Estimation of evolutionary changes in certain homologous polypeptide changes," *J. Molec. Biol.*, Vol. 64 (1972), pp. 163–179.