J. Mol. Evol. 16, 211–267 (1980)

Journal of Molecular Evolution

Theoretical Foundations for Quantitative Paleogenetics

PART III: The Molecular Divergence of Nucleic Acids and Proteins for the Case of Genetic Events of Unequal Probability

Richard Holmquist[1] and Dennis Pearl[2]

[1] Space Sciences Laboratory, University of California at Berkeley, Berkeley, California 94720, USA

[2] Department of Statistics, Graduate Division, University of California at Berkeley, Berkeley, California, 94720, USA

Summary. REH theory is extended by deriving the theoretical equations that permit one to analyze the nonrandom molecular divergence of homologous genes and proteins. The nonrandomicities considered are amino acid and base composition, the frequencies with which each of the four nucleotides is replaced by one of the other three, unequal usage of degenerate codons, distribution of fixed base replacements at the three nucleotide positions within codons, and distributions of fixed base replacements among codons. The latter two distributions turn out to dominate the accuracy of genetic distance estimates. The negative binomial density is used to allow for the unequal mutability of different codon sites, and the implications of its two limiting forms, the Poisson and geometric distributions, are considered. It is shown that the fixation intensity – the average number of base replacements per variable codon – is expressible as the simple product of two factors, the first describing the asymmetry of the distribution of base replacements over the gene and the second defining the ratio of the average probability that a codon will fix a mutation to the probability that it will not. Tables are given relating these features to experimentally observable quantities in α hemoglobin, β hemoglobin, myoglobin, cytochrome c, and the parvalbumin group of proteins and to the structure of their corresponding genes or mRNAs. The principal results are (1) more accurate methods of estimating parameters of evolutionary interest from experimental gene and protein sequence data, and (2) the fact that change in gene and protein structure has been a much less efficient process than previously believed in the sense of requiring many more base replacements to effect a given structural change than earlier estimation procedures had indicated. This inefficiency is directly traceable to Darwinian selection for the nonrandom gene or protein structures necessary for biological function. The application of these methods is illustrated by detailed consideration of the rabbit α- and β hemoglobin mRNAs and the proteins for which they code. It is found that these two genes are separated by about 425 fixed base replacements, which is a factor of two

0022–2844/80/0016/0211/$ 11.40

greater than earlier estimates. The replacements are distributed over approximately 114 codon sites that were free to accept base mutations during the divergence of these two genes.[1]

Key words: Nucleic acids – Proteins – Natural selection – Genetics – Nonrandom molecular divergence – Nonrandom REH theory – Evolution – mRNA – DNA

Introduction

The purpose of this paper is to derive the equations of nonrandom molecular divergence with respect to gene (or mRNA) and protein structure. The theoretical results are given in the form of tabular values which relate the consequences of nonrandom phenomena to observable experimental quantities for five gene families: α hemoglobin, β hemoglobin, myoglobin, cytochrome c, and the parvalbumin group of genes. These results are compared to the random theory of molecular divergence proposed by Holmquist, Cantor, and Jukes (1972) and Jukes and Holmquist (1972) and subsequently developed by them (Holmquist 1976a; Moore et al. 1976; Holmquist et al. 1976; Holmquist 1978a; Holmquist 1980). It should be noted in passing that both the random and nonrandom theory allow for the effect of Darwinian selection in restricting the number of nucleotide loci within the gene, or amino acid loci within the protein, that may fix mutations.

There are five types of nonrandom effects to be considered. First, the nucleotide composition of structural genes is usually nonrandom: the ratios A:C:G:T are not 1:1:1:1. Second, the likelihoods of a given nucleotide mutating to and being fixed as another nucleotide during evolutionary divergence are not equiprobable: the twelve transition[2] types, A → C for example, do not all occur equally frequently. Third, the probability that each of the three nucleotide positions within a codon sustains a fixed mutation may differ for each position: the third coding position usually fixes more mutations that either of the first two positions because of the extensive degeneracy of the genetic code at the third position. Fourth, codon usage, that is the frequency with which a given amino acid is coded for by a given codon, will not in general be that of the genetic code table in which each codon is used equally frequently. Fifth, the distribution of fixed mutations among codons may be uneven: the assumption this distribution is approximated by the Poisson density is not always the best (Fitch and Markowitz 1970; Uzzell and Corbin 1971).

---

[1] In common parlance the word *random* has the meaning of events of equal probability, and we shall accordingly use the word *nonrandom* for events of unequal probability. We point out, however, that in the field of statistics the word *random* is used in the more general sense of a variable capable of assuming one or more values, not necessarily of equal probability, each value a possible outcome of an experiment. A statistician might prefer the word *uniform* to *random*

[2] Throughout this paper the term *transition* is used in its general sense of an interchange, not in its specialized genetic sense restricted only to purine ↔ purine or pyrimidine ↔ pyrimidine interchanges

Although all investigators are in agreement on the existence and biological importance of these various nonrandomicities, there has not been any quantitative theory to utilize this information to increase the understanding of evolutionary phenomena. Do these effects affect the magnitude of observable consequences to change them much from the answers given by random theory? One might guess not, because both stochastic random theory and the method of maximum parsimony, which is decidedly nonrandom in its approach, estimate similar numbers for the total mutations fixed during the evolutionary descent of two genes from a common shared ancestral gene (Moore et al. 1976; Holmquist et al. 1976). However, the assumptions of both stochastic random theory and the parsimony method stray from biological reality in obvious, but different ways (Holmquist 1976a,b; Holmquist 1979). The methods given in the present paper permit a more accurate estimation of the total mutations fixed and hence also of calculations of rates of molecular divergence.

2. Fundamental Parameters During Gene Change: Definition and Assignment of Numerical Values

Gene behavior is not constant in time. It can change because of intrinsic changes in the mutation rate or because of environmental changes affecting the process of natural selection which determines whether or not a mutation will be fixed in the population. That portion of a gene which is able to accept mutations changes as successive mutations occur (Karon 1979). The experimentally observed approximate linearity in evolutionary rates over longer time periods (Fitch 1976), the stability of amino acid or gene compositions about their mean values within a family of given biological functionality, (Holmquist and Cimino 1980), the small standard deviations of the fixation intensity and number of varions for a given taxon (Moore et al. 1976; Holmquist et al. 1976) all attest that these changes occur slowly and in such a manner that though a given evolutionary parameter may wander about its mean value, it does not over longer time periods monotonically wander away from it: these wanderings are ultimately constrained by the requirements of biological function. Thus in all that follows, the basic evolutionary parameters are taken to be average values and stable. This permits us to maintain most of the mathematical simplicity of a random model while still allowing for large nonrandomicities when they are present.

We shall need to use several parameters important to genic change in our derivations and the present subsection addresses itself to the definition of these parameters and the assignment of numerical values to them.

2.1 Gene or mRNA Base Composition. The experimentally found average mole fraction of A (adenosine), C (cytidine), G (guanosine), and U (uridine) for the varied codon loci of α hemoglobin, β hemoglobin, myoglobin, cytochromes *c*, and the parvalbumin group of genes was taken from the tabulation of these values in Holmquist and Cimino 1980. The gene or mRNA base composition will be designated B_{im}, the index *i* running from 1 to 4 to designate the base (A, C, G, or T(U)) and the index *m* running from 1 to 3 to designate the position within the codon. The B_{im} are constrained by the condition

$$\sum_{i=1}^{4} B_{im} = 1.$$

2.2 Nucleotide Transition Probabilities. These are not measurable experimentally because of the inaccessibility of the nucleic acid and protein sequences ancestral to contemporary forms. At first we considered using the transition probabilities inferred by the maximum parsimony principle and reported by Goodman and Moore (1977). However, doubts about the accuracy of the reconstructed ancestral sequences (Peacock and Boulter 1975; Schwartz and Dayhoff 1978; Barker et al. 1978; Holmquist 1978b, c; Holmquist 1979), even among those who use the method, and our recent finding that, if stable, the reported maximum parsimony transition probabilities require the base composition to eventually wander *away* from the experimentally measured base composition led us in the end to employ less biased transition probabilities that maintained the experimentally observed compositional fidelity (Holmquist and Cimino 1980). With respect to notation, if a gene locus is occupied by the base B and if it changes, and is replaced in one-step by the base B' $B \rightarrow B'$ ($A \rightarrow C$, for example) the probability of this event will be designated

$$p'_{BB'} \equiv \text{Prob}(B \rightarrow B' \mid B), \quad \sum_{B' \neq B} p'_{BB'} = 1.$$

These conditional transition probabilities can be directly calculated from the asymmetric unconditional transition probabilities tabulated in Holmquist and Cimino (1980): thus

$$p'_{AC} = \frac{p_{AC}}{p_{AC} + p_{AG} + p_{AU}},$$

for example, where p_{AC}, etc. is the proportion of all one-step transitions of the type $A \rightarrow C$. The conditional transition probabilities are needed below to calculate the probability of back mutation.

2.3. Probability of Back Mutation. At a given nucleotide position m(= 1 to 3) within the codon, the probability ${}_m p_{BB}^{(X)}$ that a base initially B(= A, C, G, or U) at that locus will after X-one step base replacements remain B is (Holmquist 1976b)

$$ {}_m p_{ii}^{(X)} = \sum_{i,j=1}^{4} {}_m \rho_{iij} \cdot {}_m s_j^X \qquad (1)$$

The index i identifies the base (A, C, G, or U) and the index j identifies the term (at most four if all the coefficients ρ_{iij} are nonzero and if all the arguments s_j are distinct). The coefficients and arguements of Eq. 1 are straightforwardly calculated (see Eq. 10 and 14 in Holmquist 1976b) from the nucleotide transition probabilities alone. Conceptually the arguments s_j are the reciprocals of the roots of the probability generating function. For each of the five gene families considered in this paper, the numerical values of these coefficients and arguments are given in the Appendix at each of the three positions within the codon. They are not given here to avoid breaking the continuity of the development. The quantitative development of these qualitative ideas is fully exposited in the Appendix.

Although the coefficients ρ and the arguments s are not independent, being calculated from the same experimental data, in an intuitive sense they can be separated. The ρ are a transformed measure of the deviation of the equilibrium base composition from randomness (equimolar amounts of each of the four bases), and the s are a measure of the rapidity with which the gene base composition will return toward the

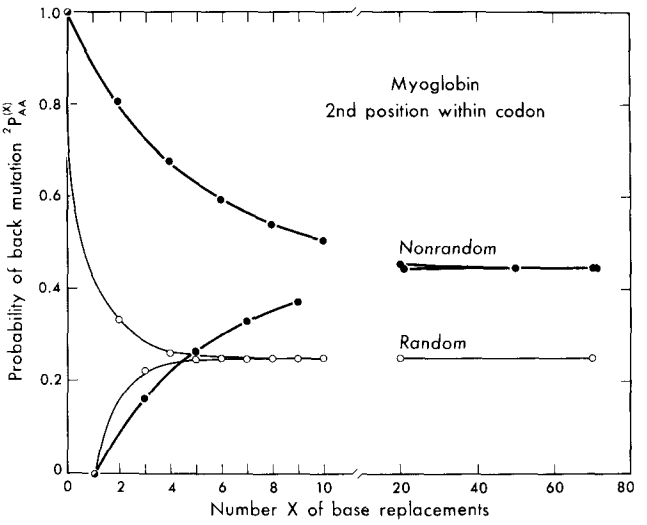

Fig. 1. Probability $^2P_{AA}^{(X)}$ for back mutation at gene loci originally adenosine at the 2nd nucleotide position within the varied codons in myoglobin. Upper curves are for X an even integer; lower curves are for X an odd integer. —— nonrandom transition probabilities; —— random transition probabilities (Calculated from Table A 4 and Eq. 1)

equilibrium values if displaced from it. This return is most rapid for a random process (Fig. 1), for which the set $[^m\rho_{iij}]$ is $[1/4, 1/4, 1/4, 1/4]$[3], independent of m and i, and the set $[^ms_j]$ is $[1, -1/3, -1/3, -1/3]$[4].

The index i identifies the base (A, C, G, or U) and the index j identifies the term (at most four if all the coefficients ρ_{iij} are nonzero and if all the arguments s_j are distinct). The coefficients and arguments of Eq. 1 are straightforwardly calculated [(see Eq. 10 and 14 in Holmquist (1976b)] from the nucleotide transition probabilities alone. Conceptually the arguments s_j are the reciprocals of the roots of the probability generating function. For each of the five gene families considered in this paper, the numerical values of these coefficients and arguments are given in the Appendix at each of the three positions within the codon. They are not given here to avoid breaking the continuity of the development.

2.4. Distribution of Fixed Mutations Within a Codon. Suppose a codon is hit n times, and designate by x_i the number of hits at the i^{th} position of the codon (i = 1 to 3). Then $x_1 + x_2 + x_3 = n$ and the joint probability distribution of x_1, x_2, and x_3 is, for

[3] It is a numerical accident that for random divergence these coefficients are equal to the molar base composition. In general they are not, though the first coefficient, for a given i, is always equal to the equilibrium composition for that base

[4] The coefficient s_1 is always unity. The fact that s_2, s_3, and s_4 are for random divergence the negative of the conditional nucleotide transition probabilities is again a numerical coincidence, not true in general

independent hits, multinomial

$$(2) \quad P(x_1, x_2, x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

$$p_1 + p_2 + p_3 = 1 \quad ,$$

where p_1, p_2, and p_3 are the probabilities that a one-step base replacement occurs at the 1st, 2nd, or 3rd position within the codon respectively. From the extensive maximum parsimony reconstructions for the five gene families considered here, the inferred base replacements reported by Goodman and Moore (1977) can be used to estimate the ratio p_1/p_2. The result is 1.23 ± 0.02 (standard deviation of the mean) irrespective of the gene family (α hemoglobin, β hemoglobin, myoglobin, cytochromes *c*, or the parvalbumin group). To what extent this concordance for different functional gene families is an artifact of the special assumptions of parsimony is not known, but it is known, experimentally, that this ratio is 1.28 if the actual numbers, 9 and 7 respectively, of base replacements observed at the first two positions within codons between the mRNAs of human and rabbit β hemoglobin are used to estimate p_1 and p_2 (Kafatos et al. 1977). In view of the fact that the main nonrandom effect in this distribution is at the 3rd position within the codon we shall simply take $p_1 = p_2$ until additional experimental nucleic acid data define this ratio better. To obtain p_3 maximum parsimony is useless because of the extensive degeneracy of the genetic code at that position. Experimentally there are 32 observed replacements at the third position within the codons for the mRNAs of rabbit and human β hemoglobin. Thus $p_3/[(p_1 + p_2)/2]$ is at least 4, and after correction for multiple hits at the same base site may be between 6 and 7. With these comments as a basis, we take

$$p_1 = 0.12$$

$$p_2 = 0.12$$

$$p_3 = 0.76.$$

For the *trpA* genes from *Salmonella typhimurium* and *Escherichia coli*, Nichols and Yanofsky (1979) reported 36, 17, and 146 observable nucletide replacements at the three codon positions. Taken at face value, that is without correcting for superimposed fixed mutations, this corresponds to $p_1:p_2:p_3::0.181:0.085:0.734$. We shall return to this matter in the *Discussion* after having developed the basis for a more exact estimation of these three parameters.[5]

[5]When the estimation is made as in the present section, it is important to choose sequences that are neither too closely nor too distantly related. In the former case there may not have been time for the differential pattern at the three positions within the codon to be realized (too few fixed mutations), and in the latter case each of the three positions will, because of mutational saturation, appear to have fixed approximately equal numbers of mutations

3. The Problem Stated and Its Solution

The consequences of accepted random gene mutations within a codon have been thoroughly explored in the REH theory for proteins published by Holmquist, Cantor and Jukes (1972), and in REH theory for genes published by Holmquist (1980). Our approach will be to obtain the analogues to Table 1, Table 5 (Table A2), and Table 8 in the former paper for accepted nonrandom point muations within a codon.

3.1 Genes or mRNA. We seek then the probabilities $P_n(\theta)$, corrected for multiple hits at the same base site, revertants, and parallelisms, that a codon hit n times has exactly θ base sites which differ from the homologous sites in the original codon (θ = 0, 1, 2, and 3).

Let $p(x_m)$ be the probability that a base at the m^{th} position within the codon (m = 1, 2, or 3) is *unchanged* after x fixed mutations. The bases at gene loci can remain unchanged in four ways: A may remain A, C may remain C, G may remain G, or T (U in mRNA) may remain T(U). $p(x_m)$ is thus the sum of four terms, each the product of the probability that the original gene locus is occupied by a given base, given by the base composition B_{im} at the varied loci, and the probability that if the locus is occupied by that base, it will remain the same after x_m fixed mutations, this latter probability being given by Equation (1). Thus,

$$p(x_m) = \sum_{i,j=1}^{4} B_{im} \cdot {}^{m}\rho_{iij} \cdot s_j^{x_m} \quad . \tag{3}$$

Then, letting angle brackets denote expectation values,

$$P_n(0) = \langle p(x_1)p(x_2)p(x_3)\rangle \quad ,$$

$$\begin{aligned} P_n(1) &= \langle [1-p(x_1)]\, p(x_2)p(x_3)\rangle + \langle [1-p(x_2)]p(x_3)p(x_1)\rangle + \\ &\quad \langle [1-p(x_3)]p(x_1)p(x_2)\rangle \\ &= \langle p(x_1)p(x_2)\rangle + \langle p(x_1)p(x_3)\rangle + \langle p(x_2)p(x_3)\rangle - 3P_n(0) \quad , \end{aligned} \tag{4}$$

$$\begin{aligned} P_n(2) &= \langle [1-p(x_1)]\,[1-p(x_2)]p(x_3)\rangle + \langle [1-p(x_2)]\,[1-p(x_3)]p(x_1)\rangle + \\ &\quad \langle [1-p(x_3)]\,[1-p(x_1)]p(x_2)\rangle \quad , \\ &= \langle p(x_1)\rangle + p\langle x_2\rangle + \langle p(x_3)\rangle - 2P_n(1) - 3P_n(0) \end{aligned}$$

$$P_n(3) = 1 - P_n(0) - P_n(1) - P_n(2) \quad .$$

Thus we need formulas for calculating only the following three expressions to complete our derivation: $\langle p(x_1)p(x_2)p(x_3)\rangle$, $\langle p(x_k)p(x_l)\rangle$, and $\langle p(x_k)\rangle$, where $k \neq l$ and k, l are 1, 2, or 3. Because of Eq. 3, this exercise reduces to calculating expectation values of the form

$$\langle s_j^{x_t} s_k^{x_u} s_l^{x_v}\rangle, \quad \langle s_j^{x_t} s_k^{x_u}\rangle \text{ and } \langle s_j^{x_t}\rangle \quad .$$

From Equations 2,

$$<\alpha^{x_t}\beta^{x_u}\gamma^{x_v}> = \sum_{x_t=0}^{n}\ \sum_{x_u=0}^{n-x_t} P(x_t,x_u,n-x_t-x_u)\,\alpha^{x_t}\beta^{x_u}\gamma^{n-x_t-x_u} = [\alpha p_t + \beta p_u + \gamma p_v]^n$$

$$<\alpha^{x_t}\beta^{x_u}> = [1 - p_t(1-\alpha) - p_u(1-\beta)]^n \qquad (5)$$

$$<\alpha^{x_t}> = [1 - p_t(1-\alpha)]^n$$

From Equations 3 and 5

$$<p(x_1)p(x_2)p(x_3)> = \sum_{\substack{i,j,k=1\\ t,u,v}}^{4} B_{i1}B_{j2}B_{k3}\,{}^1\rho_{iit}\,{}^2\rho_{jju}\,{}^3\rho_{kkv} \cdot ({p_1}^1 s_t + {p_2}^2 s_u + {p_3}^3 s_v)^n$$

$$<p(x_k)p(x_l)> = \sum_{\substack{i,j=1\\ t,u}}^{4} B_{ik}B_{jl}\,{}^k\rho_{iit}\,{}^l\rho_{jju} \cdot [1 - p_k(1 - {}^k s_t) - p_l(1 - {}^l s_u)]^n \qquad (6)$$

$$<p(x_k)> = \sum_{i,t=1}^{4} B_{ik}\,{}^k\rho_{iit}[1 - p_k(1 - {}^k s_t)]^n \quad .$$

There are a minimum of 4,096 terms to evaluate in the expression for $<p(x_1)p(x_2)p(x_3)>$, and if the coefficients ρ and arguments s are complex, one must calculate at least 65,536 terms. Even for $<p(x_k)>$, at least 16 terms must be completed. This arithmetic is trivial and inexpensive on a computer, and we will supply the Fortran program to interested investigators.

Equations 6 when substituted in Eq. 4 complete the solution to nonrandom divergence within a codon.

To return to fundamentals, the final expressions for $P_n(\theta)$ the probabilities that a codon fixing *n* mutations has exactly θ base sites which differ from the original codon- require only the base composition (three independent parameters), the conditional base transition probabilities (eight independent parameters), the distribution of fixed mutations within the codon (two independent parameters), and the total number n of mutations fixed within the codon (one independent parameter). To describe nonrandom genetic divergence adequately thus requires a minimum of 14 parameters, and if the distribution of fixed mutations among the varied codon loci is not Poison, will require more. By contrast the original random REH model (Holmquist, Cantor and Jukes 1972) was a 1-parameter model requiring only the total number n of mutations fixed within the codon. It strikes us as remarkably fortunate that the detailed consideration of the various nonrandomicities accounted for here complicates the theoretical structure so little that the observable consequences of these nonrandom effects can still be calculated with relative ease.

3.2. Proteins. Although experimental mRNA or DNA nucleotide sequence data is becoming increasingly available, much of the published experimental data is in the form of the amino acid sequences of proteins (Dayhoff 1972; Croft 1973). In the latter, molecular divergence is evidenced by homologous amino acid replacements of specific types between contemporary proteins. It is convenient to classify these types into four observational groups: those replacements of the mimimal 1-base type, minimal 2-base type, minimal 3-base type; those homologous amino acid loci that are the same in two contemporary proteins are placed in the group of minimal 0-base type. The reasons for this choice of classification have been given elsewhere (Holmquist et al. 1972; Holmquist 1976a and 1978a) and are not repeated here. The probability $P_n(\delta)$ that an amino acid replacement will be classified observationally as a minimal δ-base type ($\delta = 0, 1, 2, 3$) is

$$P_n(\delta) = \sum_{\theta=0}^{3} P_n(\theta) P_\theta(\delta) , \tag{7}$$

where the $P_n(\theta)$ are given by Eqs. 4 and $P_\theta(\delta)$ is the fraction of actual θ-base chances that will be classified as a minimal δ-base change. For α hemoglobin, β hemoglobin, myoglobin, cytochromes *c* and the parvalbumin group, the information necessary to caclulate $P_\theta(\delta)$ is given in the upper sections of Tables 1-8, respectively, and the actual values of $P_\theta(\delta)$ are given in the lower part of those tables.

The entries in Tables 1-8 were derived as follows. There are a total of 576 ($= 4^3 \times 3 \times 3$) directed single base replacements between codons. The fraction of these of a given type, say $f_{AGC/AAC}$ for the replacement of G by A at the second position within the AGC codon is given by,

$$f_{\substack{AGC \\ AAC}} = 576\, P_{AGC} p_2 {}^2p'_{GA} . \tag{8}$$

Here, P_{AGC} is the expected frequency of occurrence of the codon AGC. This may be known experimentally if mRNA or DNA sequence data is available for some genes of a family (Kafatos et al. 1977). Otherwise it may be estimated from the base composition to the extent the composition at each of the three positions within the codon are independent of each other:

$$P_{AGC} = B_{11} B_{32} B_{23} . \tag{9}$$

In Eq. 8, p_2 is the probability of a replacement occurring at the second position within the codon; ${}^2p'_{GA}$ the conditional probability that is the second position within a codon is occupied by G, and it changes, it will change to A (rather than to C or T(U); and B_{im} is the mole fraction of the base i(i = 1, 2, 3, or 4: 1 = A, 2 = C, 3 = G 4 = T(U)) at the m^{th} nucleotide position within the codon. The expected frequency of the remaining 575 interchanges is calculated similarly.

There are a total of 1728 ($4^3 \times 3 \times 3 \times 3$) directed two-base replacements between codons, and the expected frequency of occurrence of each, say $f_{AGC/AAA}$, is calculated analogously to

$$f_{\substack{AGC \\ AAA}} = 1728\, P_{AGC} \frac{p_2 p_3}{p_1 p_2 + p_1 p_3 + p_2 p_3} {}^2p'_{GA} {}^3p'_{CA} . \tag{10}$$

Table 1. Both transitions and transversions permitted: nonrandom base-change relationships in codon interchanges for α hemoglobin ($p_1 = 0.12$, $p_2 = 0.12$, $p_3 = 0.76$)

Type of change	Example	Number	
Single base replacements			
Silent		336.7	
Term-Term	UAA-UGA		3.5
Term-Amino acid	UAG-UAU		25.9
Degenerate	GGA-GGC		307.3
Recognizable as such	UCU-UAU	239.3	
Two base replacements			
Silent		110.5	
Term-Term	UAG-UGA		0.8
Term-Amino acid	UAA-CAG		97.3
Degenerate	UUA-CUC		12.4
Recorded as single base changes	UGC-AGA	1340.3	
Recognizable as such	UGC-GUC	277.1	
Three base replacements			
Silent		102.0	
Term-Term			0.
Term-Amino acid	UAG-AUU		99.0
Degenerate	UCU-AGC		3.0
Recorded as single base changes	UUA-CCC	129.0	
Recorded as two base changes	CUU-AAG	1448.3	
Recognizable as such	AUG-GAC	48.6	

Both transitions and transversions permitted: nonrandom probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.562219	0.437781	0.	0.
2	0.007616	0.822347	0.170038	0.
3	0.001869	0.079214	0.889075	0.029842

This table excludes homologous codon pairs at least one member of which is a terminating codon.

The average base composition for the first position was <A> = 0.2349, <C> = 0.2169, <G> = 0.3798 and <U> = 0.1684. In the second position it was <A> = 0.2892, <C> = 0.3229, <G> = 0.1240 and <U> = 0.2639. In the third position it was <A> = 0.2263, <C> = 0.2647, <G> = 0.2443 and <U> = 0.2647. The conditional nucleotide transition probabilities for the first position were A → C, G, U = 0.1833, 0.6721, 0.1446; C → A, G, U = 0.2146, 0.6210, 0.1644; G → A, C, U = 0.3868, 0.3531, 0.2601; and U → A, C, G = 0.2462, 0.2359, 0.5180. In the second position they were A → C, G, U = 0.5226, 0.1360, 0.3414; C → A,G, U = 0.4629, 0.1443, 0.3928; G → A, C, U = 0.3274, 0.3635, 0.3091; and U → A, C, G = 0.3756, 0.4801, 0.1444. In the third position, they were A → C, G, U = 0.3454, 0.3091, 0.3454; C → A, G, U = 0.2954, 0.3293, 0.3752, G → A, C, U = 0.2862, 0.3569, 0.3569; and U → A, C, G = 0.2954, 0.3752, 0.3293. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.12, 0.12 and 0.76 respectively.

Codon usage was estimated from the above average nucleotide composition at the varied codon acid loci of 51 α hemoglobin chains on the assumption that the three positions within the codon were independentas per Eqs. 8, 9, 10 and 11 in text

Table 2. Both transitions and transversions permitted: nonrandom base-change relationships in codon interchanges for α hemoglobin ($p_1 = p_2 = p_3 = 1/3$)

Type of change	Example	Number	
Single base replacements			
Silent		168.2	
Term-Term	UAA-UGA		2.0
Term-Amino acid	UAG-UAU		28.9
Degenerate	GGA-GGC		137.2
Recognizable as such	UCU-UAU	407.8	
Two base replacements			
Silent		108.1	
Term-Term	UAG-UGA		0.6
Term-Amino acid	UAA-CAG		97.8
Degenerate	UUA-CUC		9.8
Recorded as single base changes	UGC-AGA	1001.3	
Recognizable as such	UGC-GUC	618.6	
Three base replacements			
Silent		102.0	
Term-Term			0.
Term-Amino acid	UAG-AUU		99.0
Degenerate	UCU-AGC		3.0
Recorded as single base changes	UUA-CCC	129.0	
Recorded as two base changes	CUU-AAG	1448.3	
Recognizable as such	AUG-GAC	48.6	

Both transitions and transversions permitted: nonrandom probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.251780	0.748220	0.	0.
2	0.006003	0.614412	0.379585	0.
3	0.001869	0.079214	0.889075	0.029842

As for Table 1 except that the relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.3333, 0.3333 and 0.3333, respectively

Table 3. Both transitions and transversions permitted: random base-change relationships in codon interchanges

Type of change	Example	Number	
Single base replacements			
Silent		184.0	
Term-Term	UAA-UGA		4.0
Term-Amino acid	UAG-UAU		46.0
Degenerate	GGA-GGC		134.0
Recognizable as such	UCU-UAU	392.0	
Two base replacements			
Silent		188.0	
Term-Term	UAG-UGA		2.0
Term-Amino acid	UAA-CAG		158.0
Degenerate	UUA-CUC		28.0
Recorded as single base changes	UGC-AGA	1006.0	
Recognizable as such	UGC-GUC	534.0	
Three base replacements			
Silent		174.0	
Term-Term			0.
Term-Amino acid	UAG-AUU		162.0
Degenerate	UCU-AGC		12.0
Recorded as single base changes	UUA-CCC	308.0	
Recorded as two base changes	CUU-AAG	1164.0	
Recognizable as such	AUG-GAC	82.0	

Both transitions and transversions permitted: random probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.254753	0.745247	0.	0.
2	0.017857	0.641582	0.340561	0.
3	0.007663	0.196679	0.743295	0.052363

This table excludes homologous codon pairs at least one member of which is a terminating codon. The average base composition at all three positions within the codon were taken to <A> = 0.2500, <C> = 0.2500, <G> = 0.2500 and <U> = 0.2500. The conditional nucleotide transition probabilities for all three positions within the codon were taken to be A → C, G, U = 1/3, 1/3, 1/3, etc. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 1/3 each

Table 4. Both transitions and transversions permitted: nonrandom base-change relationships in codon interchanges for β hemoglobin

Type of change	Example	Number	
Single base replacements			
Silent		314.4	
Term-Term	UAA-UGA		0.7
Term-Amino acid	UAG-UAU		7.9
Degenerate	GGA-GGC		305.7
Recognizable as such	UCU-UAU	258.0	
Two base replacements			
Silent		45.3	
Term-Term	UAG-UGA		0.4
Term-Amino Acid	UAA-CAG		43.3
Degenerate	UUA-CUC		1.6
Recorded as single base changes	UGC-AGA	1334.6	
Recognizable as such	UGC-GUC	337.2	
Three base replacements			
Silent		39.6	
Term-Term			0.
Term-Amino acid	UAG-AUU		37.9
Degenerate	UCU-AGC		1.7
Recorded as single base changes	UUA-CCC	65.9	
Recorded as two base changes	CUU-AAG	1544.3	
Recognizable as such	AUG-GAC	67.4	

Both transitions and transversions permitted: nonrandom probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

Θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.542311	0.457689	0.	0.
2	0.000948	0.797527	0.201525	0.
3	0.001012	0.039228	0.919615	0.040145

This table excludes homologous codon pairs at least one member of which is a terminating codon.

The average nucleotide composition for the first position was <A> = 0.2238, <C> = 0.2069, <G> = 0.4506 and <U> = 0.1187. In the second position it was <A> = 0.3235, <C> = 0.2311, <G> = 0.1471 and <U> = 0.2983. In the third position it was <A> = 0.0683, <C> = 0.2784, <G> = 0.3581 and <U> = 0.2952.

The conditional nucleotide transition probabilities for the first position were A → C, G, U = 0.0384, 0.9397, 0.0219; C → A, G, U = 0.0852, 0.8604, 0.0544; G → A, C, U, = 0.3946, 0.3799, 0.2276; and U → A, C, G = 0.2391, 0.2363, 0.5246. In the second position they were A → C, G, U = 0.3224, 0.1772, 0.5004; C → A, G, U = 0.4454, 0.1806, 0.3740; G → A, C, U = 0.3777, 0.2823, 0.3399; and U → A, C, G = 0.5531, 0.2858, 0.1610. In the third position they were A → C, G, U = 0.3197, 0.3558, 0.3245; C → A, G, U = 0.0793, 0.5650, 0.3558; G → A, C, U = 0.0717, 0.4424, 0.4859; and U → A, C, G = 0.0696, 0.3324, 0.5979. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.12, 0.12 and 0.76, respectively.

The frequency of a given interchange was based on the actual codon usage (Kafatos et al. 1977) in human (Marotta et al. 1977) and rabbit (Efstratiadis et al. 1977) beta hemoglobin messenger RNA

Table 5. Both transitions and transversions permitted: nonrandom base-change relationships in codon interchanges for β hemoglobin

Type of change	Example	Number	
Single base replacements			
Silent		326.8	
Term-Term	UAA-UGA		1.1
Term-Amino acid	UAG-UAU		19.9
Degenerate	GGA-GGC		305.8
Recognizable as such	UCU-UAU	249.2	
Two base replacements			
Silent		70.8	
Term-Term	UAG-UGA		0.3
Term-Amino acid	UAA-CAG		65.3
Degenerate	UUA-CUC		5.2
Recorded as single base changes	UGC-AGA	1347.1	
Recognizable as such	UGC-GUC	310.1	
Three base replacements			
Silent		67.5	
Term-Term			0.
Term-Amino acid	UAG-AUU		66.0
Degenerate	UCU-AGC		1.5
Recorded as single base changes	UUA-CCC	75.6	
Recorded as two base changes	CUU-AAG	1504.7	
Recognizable as such	AUG-GAC	80.2	

Both transitions and transversions permitted: nonrandom probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.550994	0.449006	0.	0.
2	0.003139	0.810324	0.186537	0.
3	0.000915	0.045459	0.905352	0.048275

This table excludes homologous codon pairs at least one member of which is a terminating codon.

The average base composition for the first position was <A> = 0.2215, <C> = 0.1794, <G> =0.4463 and <U> = 0.1528. In the second position it as <A>= 0.3297, <C>= 0.2174, <G> = 0.1551 and <U> = 0.2978. In the third position it was <A> = 0.0683, <C> = 0.2784, <G> = 0.3581 and <U> = 0.2952. The conditional nucleotide transition probabilities for the first position were A → C, G, U = 0.0205, 0.9623, 0.0172; C → A, G, U = 0.1326, 0.7543, 0.1131; G → A, C, U = 0.3797, 0.3319, 0.2884; and U → A, C, G = 0.1848, 0.1748, 0.6403. In the second position they were A → C, G, U = 0.3014, 0.1925, 0.5062; C → A, G, U = 0.4462, 0.1933, 0.3606; G → A, C, U = 0.3919, 0.2694, 0.3387; and U → A, C, G = 0.5773, 0.2560, 0.1666. In the third position they were A → C, G, U = 0.3197, 0.3558, 0.3245; C → A, G, U = 0.0793; 0.5650, 0.3558; G → A, C, U = 0.0717, 0.4424, 0.4859; and U → A, C, G = 0.0696, 0.3324, 0.5979. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.12, 0.12 and 0.76, respectively.

Codon usage was estimated from the average nucleotide composition given above at the varied codon loci of 59 β hemoglobin chains on the assumption that the three positions within the codon behaved independently as per Eqs. 8, 9, 10, and 11 in text

Table 6. Both transitions and transversions permitted: nonrandom base-change relationships in codon interchanges for myoglobin

Type of change	Example	Number	
Single base replacements			
Silent		293.9	
Term-Term	UAA-UGA		3.4
Term-Amino Acid	UAG-UAU		21.2
Degenerate	GGA-GGC		269.2
Recognizable as such	UCU-UAU	282.1	
Two base replacements			
Silent		93.7	
Term-Term	UAG-UGA		1.1
Term-Amino acid	UAA-CAG		81.9
Degenerate	UUA-CUC		10.7
Recorded as single base changes	UGC-AGA	1244.7	
Recognizable as such	UGC-GUC	389.6	
Three base replacements			
Silent		84.7	
Term-Term			0.
Term-Amino acid	UAG-AUU		84.1
Degenerate	UCU-AGC		0.6
Recorded as single base changes	UUA-CCC	67.4	
Recorded as two base changes	CUU-AAG	1517.3	
Recognizable as such	AUG-GAC	58.7	

Both transitions and transversions permitted: nonrandom probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.488290	0.511710	0.	0.
2	0.006505	0.756635	0.236860	0.
3	0.000354	0.040990	0.922945	0.035712

This table excludes homologous codon pairs at least one member of which is a terminating codon.

The average base composition for the first position was <A> = 0.3070, <C> = 0.2092, <G> = 0.3911 and <U> = 0.0927. In the second position it was <A> = 0.4477, <C> = 0.2181, <G> = 0.1223 and <U> = 0.2119. In the third position it was <A> = 0.2857, <C> = 0.2224, <G> = 0.2695 and <U> = 0.2224. The conditional nucleotide transition probabilities for the first position were A → C, G, U = 0.1601, 0.7883, 0.0516; C → A, G, U = 0.3171, 0.5388, 0.1441; G → A, C, U = 0.5399, 0.3406, 0.1194; and U → A, C, G = 0.3182, 0.2893, 0.3925. In the second position they were A → C, G, U = 0.3863, 0.2326, 0.3811; C → A, G, U = 0.9073, 0.0353, 0.0574; G → A, C, U = 0.5286, 0.2361, 0.2352; and U → A, C, G = 0.8738, 0.0768, 0.0494. In the third position they were A → C, G, U = 0.2977, 0.4047, 0.2977; C → A, G, U = 0.3816, 0.3460, 0.2725; G → A, C, U = 0.4303, 0.2848, 0.2848; and U → A, C, G = 0.3816, 0.2725, 0.3460. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.12, and 0.76, respectively.

Codon usage was estimated from the average nucleotide composition given above at the varied codon loci of 44 myoglobin chains on the assumption that the three positions within the codon behaved independently as per Eqs. 8, 9. 10, and 11 in text

Table 7. Both transitions and transversions permitted: nonrandom base-change relationships in codon interchanges for cytochrome *c*

Type of change	Example	Number	
Single base replacements			
Silent		302.4	
Term-Term	UAA-UGA		4.9
Term-Amino acid	UAG-UAU		30.4
Degenerate	GGA-GGC		267.1
Recognizable as such	UCU-UAU	273.6	
Two base replacements			
Silent		128.7	
Term-Term	UAG-UGA		1.2
Term-Amino acid	UAA-CAG		118.1
Degenerate	UUA-CUC		9.4
Recorded as single base changes	UGC-AGA	1230.5	
Recognizable as such	UGC-GUC	368.8	
Three base replacements			
Silent		122.0	
Term-Term			0.
Term-Amino acid	UAG-AUU		120.4
Degenerate	UCU-AGC		1.6
Recorded as single base changes	UUA-CCC	83.4	
Recorded as two base changes	CUU-AGG	1452.9	
Recognizable as such	AUG-GAC	69.7	

Both transitions and transversions permitted: nonrandom probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.493926	0.506074	0.	0.
2	0.005868	0.764862	0.229270	0.
3	0.000999	0.051881	0.903768	0.043352

This table excludes homologous codon pairs at least one member of which is a terminating codon. The average base composition for the first position was <A> = 0.3687, <C> = 0.1266, <G> = 0.3631 and <U> = 0.1416. In the second position it was <A> = 0.4321, <C> = 0.2626, <G> = 0.1034 and <U> = 0.2018. In the third position it was <A> = 0.2707, <C> = 0.2340, <G> = 0.2614 and <U> = 0.2340. The conditional nucleotide transition probabilities for the first position were A → C, G, U = 0.1329, 0.7114, 0.1557; C → A, G, U = 0.3783, 0.3712, 0.2505; G → A, C, U = 0.7325, 0.1230, 0.1445; and U → A, C, G = 0.3878, 0.2326, 0.3796. In the second position they were A → C, G, U = 0.4482, 0.1751, 0.3767; C → A, G, U = 0.9401, 0.0160, 0.0439; G → A, C, U = 0.4567, 0.2771, 0.2662; and U → A, C, G = 0.6838, 0.1995, 0.1167. In the third position they were A → C, G, U = 0.3134, 0.3732, 0.3134; C → A, G, U = 0.3626, 0.3427, 0.2947; G → A, C, U = 0.3867, 0.3066, 0.3066; and U → A, C, G = 0.3626, 0.2947, 0.3427. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.12, 0.12 and 0.76, respectively.

Codon usage was estimated from the average nucleotide composition given above at the varied codon loci in 57 cytochrome *c* chains on the assumption that the three positions within the codon behaved independently as per Eqs. 8, 9. 10, and 11 in text

Table 8. Both transitions and transversions permitted: nonrandom base-change relationships in codon interchanges for the parvalbumin group genes

Type of change	Example	Number	
Single base replacements			
Silent		309.4	
Term-Term	UAA-UGA		4.3
Term-Amino acid	UAG-UAU		29.2
Degenerate	GGA-GGC		276.0
Recognizable as such	UCU-UAU	266.6	
Two base replacements			
Silent		117.7	
Term-Term	UAG-UGA		1.2
Term-Amino acid	UAA-CAG		110.8
Degenerate	UUA-CUC		5.7
Recorded as single base changes	UGC-AGA	1263.7	
Recognizable as such	UGC-GUC	346.6	
Three base replacements			
Silent		113.9	
Term-Term			0.
Term-Amino acid	UAG-AUU		113.2
Degenerate	UCU-AGC		0.7
Recorded as single base changes	UUA-CCC	59.8	
Recorded as two base changes	CUU-AAG	1473.2	
Recognizable as such	AUG-GAC	81.1	

Both transitions and transversions permitted: nonrandom probability $P_\theta(\delta)$ that an actual θ-base change will be recorded as a δ-base change

Θ	P(0)	P(1)	P(2)	P(3)
0	1.000000	0.	0.	0.
1	0.508629	0.491371	0	0.
2	0.003550	0.781947	0.214503	0.
3	0.000457	0.037023	0.912276	0.050243

This table excludes homologous codon pairs at least one member of which is a terminating codon, The average base composition for the first position was <A> = 0.2833, <C> = 0.1143, <G> = 0.4549 and <U> = 0.1475. In the second position it was <A> = 0.3827, <C> = 0.2074, <G> = 0.1298 and <U> = 0.2801. In the third position it was <A> = 0.2540, <C> = 0.2400, <G> = 0.2660 and <U> = 0.2400. The conditional nucleotide transition probabilities for the first position were A → C, G, U = 0.0005, 0.9987, 0.0008; C → A, G, U = 0.2268, 0.5809, 0.1923; G → A, C, U = 0.5115, 0.2131, 0.2754; and U → A, C, G = 0.1675, 0.1168, 0.7157. In the second position they were A → C, G, U = 0.3353, 0.1857, 0.4790; C → A, G, U = 0.5559, 0.1634, 0.2807; G → A, C, U = 0.4383, 0.2646, 0.2971; and U → A, C, G = 0.7516, 0.1598, 0.0887. In the third position they were A → C, G, U = 0.3151, 0.3699, 0.3151; C → A, G, U = 0.3335, 0.3584, 0.3081; G → A, C, U = 0.3531, 0.3234, 0.3234; and U → A, C, G = 0.3335, 0.3081, 0.3584. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.12, 0.12 and 0.76, respectively.

Codon usage was estimated from the average nucleotide composition given above at the varied codon loci in 22 parvalbumin group chains on the assumption that the three positions within the codon behaved independently as per Eqs. 8, 9. 10, and 11 in text

For three-base interchanges between codons, the calculation of the expected frequency of each, say $f_{AGC/GAA}$, is analogous to

$$f_{\substack{AGC \\ GAA}} = 1728\, P_{AGC}\, {}^{1}p'_{AG}\, {}^{2}p'_{GA}\, {}^{3}p'_{CA} \quad . \tag{11}$$

3.3. Distribution of Fixed Mutations Among Codons. It is necessary to have a quantitative description of the way nucleotide point mutations that have been fixed by natural selection or neutral drift are distributed over the gene. The number of such fixations separating one gene from a second homologous gene has been designated *random evolutionary hits*, REH, by Jukes and Holmquist (1972) in recognition of the probabilistic nature of the process of base replacement and subsequent fixation. Some functionally important amino acid sites may not fix any mutations. The heme group of all known cytochromes *c* is attached to the protein chain by a cysteine at residue 17. Let us call the number of such invariant residues T_1, corresponding to $L_1 = 3T_1$ nucleotides. The simplest statement about these T_1 amino acid sites is that they do not contribute to REH at all. This would require the corresponding codons to be invariant for all species examined. Because of the degeneracy of the genetic code, this last requirement, though the most parsimonious, may not be factual. For example, some of the invariant cysteines at residue 17 in the cytochromes *c* may be coded for by UGU, others UGC. In the mRNAs for rabbit and human β hemoglobin, although amino acid residues 36, 42, 66, 82, and 145 are invariant, each is known from the experimentally determined mRNA sequences to have sustained at least one base replacement at the third position within the codon for each of these residues. If from the gene or mRNA data from several species, particular codon sites are known to be truly invariant then T_1 is the number of such sites. Let REH_1 be the total number of fixations separating the two homologous genes for the invariant region. We now direct our attention to the distribution of REH_2 fixed base replacements among the $T_2 = T - T_1$ variable codons or the variable amino acid residues, where $REH = REH_1 + REH_2$, and T is the total number of codons in the gene.

If every codon fixing mutations in the gene underwent exactly *n* base replacements, and the value of *n* is known, then the expectation values given by Eqs. 4 and Eqs. 7 could be compared directly with experiment. Of course all codons do not fix the same number of mutations. Some may fix more than the average; others may fix less; and certainly the value of *n* is not known for each separate codon. The estimation of the distribution of base replacements among codons has not to date been solved in a totally satisfactory manner. One may assume that each variable codon site has an equal probability of fixing a point nucleotide mutation so that the fraction of the variable codon sites that sustained *n* fixations is given by the binomial distribution

$$Pr(n) = \binom{REH_2}{n} \left(\frac{1}{T_2}\right)^{n} \left(1 - \frac{1}{T_2}\right)^{REH_2 - n} \quad , \tag{9}$$

which can be approximated by the Poisson density.

$$Pr(n) = \frac{e^{-\mu}\mu^{n}}{n!} \quad , \tag{10}$$

where the *fixation intensity* $\mu = REH_2/T_2$ is the average total number of base replacements per varion (variable codon sites, T_2 in number). This was the approach taken in the original development of the REH theory of molecular divergence (Holmquist et al. 1972; Jukes and Holmquist 1972).

Or one may divide the variable codon sites into two groups, one more variable than the other, each being described by its own Poisson parameter. This is the approach taken by Fitch and Markowitz (1970). The latter approach can be more accurate, but as recognized by Fitch and Markowitz, is to some extent artificial as there are surely more than only two mutability groups of codons.

The necessary revision to our model to more accurately reflect these nonuniform effects of natural selection can be approached as follows. If there are several (k for concreteness) classes of variable codon sites, the i^{th} class being $T_{(i)}$ amino acid residues in length with a mutability described by its Poisson parameter $\mu_{(i)}$, we may write

$$REH_2 = \mu_{(1)}T_{(1)} + \mu_{(2)}T_{(2)} + \cdots \mu_{(k)}T_{(k)} \tag{11}$$

Passing for convenience to the continuous limit, the proportion of variable codons receiving n fixation is then given by

$$Pr(n) = \int_0^\infty \frac{e^{-\mu}\mu^n}{n!}\, g(\mu)\, d\mu \quad , \tag{12}$$

where $g(\mu)d\mu$ is the probability that the Poisson parameter lies between the values μ and $\mu + d\mu$. It follows immediately from Eq. 11 that

$$REH_2 = T_2 \int_0^\infty \mu\, g(\mu) d\mu \quad . \tag{13}$$

We must now determine $g(\mu)$.

Uzzell and Corbin (1971) found, for cytochrome *c*, that the proportion of variable codons receiving n fixations was reasonably represented by the negative binomial density

$$Pr(n) = \binom{r + n - 1}{n} q^r p^n \quad . \tag{14}$$

Here p is the average probability that a mutation will be fixed and q (= 1 - p) the average probability that it will not, where the usual symbol in parentheses has been used for the generalized (that is r need not be integral) binomial coefficients (Feller, 1968). Equations 12 and 14 imply (Feller 1971a) that $g(\mu)$ has the gamma distribution

$$g(\mu) = \left(\frac{q}{p}\right) \frac{1}{\Gamma(r)} \left(\frac{q\mu}{p}\right)^{r-1} e^{-(q\mu/p)} \tag{15}$$

From a biological point of view, it is not obvious that one would expect the fixation intensities μ to follow this rather peculiar form proposed by Uzzell and Corbin. Their proposal becomes clearer if one analyzes the special case where r is integral in Eq. 15. Consider μ as the resultant of r mutually independent random (sources of mutability):

$$\mu = {}^1\mu + {}^2\mu + \ldots + {}^r\mu \quad , \tag{16}$$

where each contributing variable ${}^k\mu$ has the exponential distribution (Feller 1971b)

$$f({}^k\mu) = (\frac{q}{p})\, e^{-(q/p)({}^k\mu)} \quad . \tag{17}$$

Under these conditions it can be proved (Feller 1971b) that μ has the above gamma distribution. Thus one interpretation of Uzzell and Corbin's model is that μ is the sum of independently contributing sources, ${}^k\mu$, in which the larger ${}^k\mu$, the less probable it is. That is, sources which result in a high number of fixed mutations are less likely to contribute to the overall number of fixations. All this is plausible on biological grounds and makes the proposal by Uzzell and Corbin understandable. Their analysis is generally accepted (King and Wilson 1975).

It is perhaps not biologically reasonable to expect each of the sources contributing to the overall μ to decay in eaxactly the same exponential manner (q/p in Eq. 17 is independent of the source k). Whatever inflexibility or error introduced by the assumption of constant q/p appears in practice to be, in part at least, offset by the freedom to vary r, a measure of the number of independent sources contributing to the sum; otherwise Uzzell and Corbin would not have observed in their Table 4 the good fit to the data that they did. The constancy of q/p is not a severe biological restriction provided one interprets this ratio as an average value over all contributing sources of base replacement.[6]

[6] The truth of Eq. 15 follows from Eqs. 12 and 14 alone, independent of whether or not the mechanism of Eqs. 16 and 17 is correct. The latter mechanism, though seemingly plausible, is but one of several whose final result is Eq. 14. In fact Pr(n) in Eq. 14 can be derived directly as the probability that exactly n fixed mutations precede the r^{th} unfixed mutation. This gives n the interpretation of a waiting time. Consistent with, but again independent of this interpretation, is the identity

$$p \equiv \frac{\langle n \rangle}{r + \langle n \rangle} \quad ,$$

where p is the probability that a mutation will be fixed, the angle brackets denoting the expectation value of n.

If $r \to \infty$ and $p \to 0$ in such a manner that $rp \to \mu$ (a constant), Eq. 14 reduces to Eq. 10, so that Eq. 14 includes the earlier Poisson models as a special case. For $r = 1$, Eq. 14 reduces to the geometric distribution $\Pr(n) = qp^n$ and Eq. 15 reduces to Eq. 17. For not too small values of μ the Poisson distribution is quasi-symmetrical and attains its maximal value near its mean $\langle n \rangle = \mu$, and Pr(n) drops off rapidly on either side of this mean. By contrast, the geometric distribution is grossly asymmetrical, being maximal at $n = 0$ and decays slowly so that large values of n occur with nonnegligible probability.

Thus the negative binomial distribution is capable of describing the qualitative features of a wide range of possible behaviors and at the same time is consistent with a variety of mechanisms giving rise to those behaviors. Its ultimate usefulness must rest with its ability to describe the data accurately

An explicit equation for REH_2 can now be written:

$$REH_2 = T_2 \int_0^\infty \mu\, g(\mu) d\mu = r\left(\frac{p}{q}\right) T_2 \quad . \tag{18}$$

The average number of fixations per codon is, as stated earlier,

$$\mu_{av} = \frac{REH_2}{T_2} = r\left(\frac{p}{q}\right) = n_{av} \quad . \tag{19}$$

where n_{av} is the mean value of n in the negative binomial distribution (Eq.14).

In the original REH model, estimation of the total number of base replacements between two genes, $REH_2 = \mu_2 T_2$, required the estimation of two parameters, μ_2 and T_2, from the data. In the present approach $REH_2 = rpT_2/q$ requires the estimation of three parameters r, p, and $T_2 (q = 1 - p)$. The proposed theory is thus minimally more complex from a computational point of view, yet allows for Darwinian selection in a substantially more accurate manner.

We must now assign numerical values to r, p, and T_2. From their parsimonious reconstruction of the cytochrome *c* phylogeny, Uzzell and Corbin inferred the minimum number of fixations that each codon site had sustained and tabulated the number of codons having received exactly n fixations. They then used Bliss and Fisher's (1953) method to find the values of r and p that best fit that data. They found, for cytochrome *c*, $r \cong 2.05$ and $p \cong 0.668$, so that the average fixation intensity was $\mu_{av} \cong 4.51$ base replacements/codon[7]. The number of fixations inferred by the method of parsimony err on the low side – an error by a factor of two or more is not uncommon (Moore et al. 1976; Holmquist et al. 1976) – unless the data base is very dense (Holmquist 1978c). These referenced papers and the data for cytochrome *c* in Fig. 4 of Fitch (1976) indicate the data base used by Uzzell and Corbin was too sparse to permit accurate inference of the number of base replacements. The numerical values just quoted for r and p should thus be regarded as tentative and may also be different for a different protein family than the cytochromes *c*. An alternative method of estimating r and p will be given in the Section 3.5.

The total average number T_2 of variable codon sites during the divergence of two homologous genes is given by (Holmquist 1978a)

$$T_2 = \frac{1 - P(0)}{1 - P_2(0)}\, T \quad , \tag{20}$$

where P(0) is the observed proportion, *over all T codons (amino residues)*, of identical codon (amino acid) sites in the two homologous genes (proteins) under

[7] The correspondences between the notation of the Bliss and Fisher (BF) paper and the present one are $r = k_{BF}$; $p = p_{BF}/(1 + p_{BF})$; $q = 1/q_{BF}$; $\mu_{av} = rp/q = m_{BF} = p_{BF} k_{BF}$

comparison and $P_2(0)$ is that proportion *of the T_2 varions* in which the two homologous codons (or amino acids) at a given site are identical. The calculation of $P_2(0)$ is given by Eq. 21 below, for genes, and by Eq. 22 below, for proteins.

3.4 *Expected Proportion of Variable Gene or Amino Acid Sites with no, one, two or three Observable Base Replacements After an Arbitrary Distribution Pr(n) of Fixed Hits per Codon During Nonrandom Molecular Divergence.* The expected proportion of variable codons fixing exactly n base replacements was designated $\Pr(n)$ in the last section. More generally we may designate this proportion by $\Pr(n|w_i)$ to indicate that $\Pr(n)$ is determined by a set of numberical parameters $w_i - r$ and p in Eq. 14, for example. The expected proportions of the variable codons that exhibit, no, one, two or three *observable* base replacements are

$$^{\Omega}P_2(\theta) = \sum_{n=0}^{n_{max}} \Pr(n)\, P_n(\theta) \quad , \tag{21}$$

where θ ranges from 0 to 3 and $P_n(\theta)$ is given by Eq. 4. The superscript Ω on the left-hand side of Eq. 21 is to remind us that $\Pr(n)$ is a function of the numerical parameter set $\Omega \equiv \{w_i\}$. The subscript "2" is to remind us that $^{\Omega}P_2(\theta)$ is a proportion of the variable codon sites and not a proportion of all codon sites of the gene (see Section 3.3) unless the two happen to coincide. Such a coincidence would be rare because of functional constraints.

The expected proportion of variable amino acid loci in two homologous proteins the structural genes for which are descendant from a common ancestral gene and that exhibit, no, one, two or three observable *minimal* base type amino replacements is

$$^{\Omega}P_2(\delta) = \sum_{n=0}^{n_{max}} \Pr(n) P_n(\delta) \quad , \tag{22}$$

with δ ranging from 0 to 3 and $P_n(\delta)$ being given by Eq. 7. This equation is the analogue to Eq. 12 in REH theory for equiprobable genetic events (Holmquist et al. 1972) and generates the nonrandom analogue to Table A5 in that reference.Equation 22 reproduces Table A5 identically if the parameter space is Poisson (Eq. 14 this paper with $\Omega \equiv \{r, p\} = \{\infty, 0\}$).

3.5 *Estimation of the w_i.* In order to implement Eq. 21 or 22, the numerical values of the parameters w_i that determine $\Pr(n)$ must be known. There are many ways of estimating these values from the data (Mood 1950), but each method uses the same principle: those numerical values of the w_i are chosen that give the best fit, in some well-defined sense, of theory to observation.

To illustrate, using the familiar chi-square measure for goodness of fit, if the expected number of identical codon sites between two homologous genes is

$$n(0) = T_1 + T_2 P_2(0) \quad , \tag{23}$$

where T_1 is the number of invariant codon sites ($T_1 = T - T_2$) and T_2 and $P_2(0)$ are given by Eqs. 20 and 21, and if the expected number of codon sites sustaining exactly one, two, or three observable (as between mRNAs) base replacements is

$$\begin{aligned} n(1) &= P_2(1)T_2 \\ n(2) &= P_2(2)T_2 \\ n(3) &= P_2(3)T_2 \ , \end{aligned} \tag{24}$$

one substitutes Eq. 14 into Eq. 21 and finds those values of r, p, and T_2 that minimize

$$\chi^2 = \sum_{i=0}^{3} \frac{[n(i) - n'(i)]^2}{n(i)} = \sum_{i=1}^{3} \frac{[n(i) - n'(i)]^2}{n(i)} \ ,$$

where $n'(i)$ are the experimentally observed number of codon sites of each type. Iterative numerical methods for performing the minimization of functionals are published in many places (Ortega and Rheinboldt 1970). Chi-square is not the only functional that can be minimized and one should try to choose the latter so that the w_i are estimated most simply and most meaningfully for a given set of data. We realize the "chi-square" function we have minimized does not have the chi-square distribution; we are using it simply as a functional measure defining fit. Another good method for estimating parameters constrained by observational data is the maximum entropy formalism [see Levine and Tribus, (1979)].

4. Results

4.1. Back Mutation: Its Relevance to Selective Neutrality and the Hypothesis of Selective-Stochastic Stability. Figure 1 in Section 2.3 illustrates the expected behavior at a gene site originally adenosine at the second nucleotide position within myoglobin codons during nonrandom molecular divergence as contrasted to the behavior during random divergence. For an even number of base replacements the probability of an observable back mutation is higher than for random mutation and remains higher for a longer period of time before approaching its equilibrium value given by the average mole fraction of adenosine at those loci. For an odd number of base replacements the probability of observing a back mutation is low initially but increases beyond random expectation under repeated replacement. Other patterns than the one illustrated are of course possible for other situations, but the generalization which emerges from all of them is that *for a gene displaced from compositional equilibrium, the approach to the stable equilibrim values is most rapid under random fixation of mutations.* Conversely, it requires fewer base replacements to stochastically displace a gene near the random composition A:C:G:T:1:1:1:1 to a nearby composition than is required for a less random starting composition. Such compositional displacements could be important for adaptive gene changes.

Were molecular divergence a selectively neutral process so that base replacements bore no biological cost we would expect most proteins to have an amino acid compo-

sition near that given by the genetic code table, but at the same time there should be a number of proteins of quite aberrant compositions arising from stochastic drift. On the other hand were protein composition a result of Darwinian selection and if each base replacement bore some finite biological cost then by the reasoning of the preceding paragraph we would expect to see protein compositions displaced from random to maintain biological function, but the magnitude of this displacement should be small so that compositional fidelity could be maintained without too great a biological cost arising from too numerous base replacements. Small but finite compositional deviations from the random for proteins is precisely what are most often experimentally observed (Holmquist 1978d). The conceptual content of Fig. 1 thus provides both a qualitative and quantitative theoretical explanation for the hypothesis of selective-stochastic stability which was originally based on empirical observation alone (Holmquist and Moise 1975). Most likely both selectively neutral and Darwinian divergence occur.

4.2. Expected Proportion of Variable Gene Codon Sites with no, one, two, or three Observable Base Replacements After n *Fixed Hits per Codon During Nonrandom Molecular Divergence.* The desired proportion $P_n(\theta)$, $\theta = 0, 1, 2, 3$ (Section 3.1) is given for α hemoglobin in Tables 9a and 10a. The former Table is for a nonrandom distribution of fixed mutations over the three nucleotide positions within each codon, the third position receiving about six times as many replacements as the other two. In Table 10a, each position within the codon is taken as equally likely to fix a mutation: this permits us to isolate the effect of nonrandom base composition and base transition probabilities (which are given in the legends to Tables 1–8) from that of a nonrandom distribution of replacements within the codon. For a purely random process, the expected values of $P_n(\theta)$ are given in Table 11a. Comparing Tables 10a and 11a for n = 4, for example, it can be seen that nonrandom base composition – the mole fractions of the four bases span the range from about 12–38% (the exact values are in Table 1) as opposed to 25% each for the random case – and nonrandom conditional base transition probabilities – spanning the range from about 0.14 to 0.67 (exact values in Table 1) as opposed to 1/3 (0.33) for the random case – do *not* change the expected number and types of base replacements much: to illustrate, for a purely random process, after a codon has fixed four such replacements an average of 19.20% (Table 11a) of such codons will exhibit an observable change at one of the three positions within each codon, whereas for the nonrandom case of α hemoglobin the expected percentage is 19.41% (Table 10a) provided the distribution of the four fixed replacements remains even over the codons. The non-randomness in compositions and transition probabilities are for the most part self-compensating: an excess over random expectation of the number of replacements of say A → C, G or U due to, say, an excess of A over 25% and/or a high probability of the transitions A → C, A → G, or A → U will be partially offset by fewer mutations from, say, a deficit of G less than 25% and/or a low probability of some of the other transitions. The compensation is, however, not exact: by examining $\langle\theta\rangle$, the average number of sites per codon which differ from the homologous sites in the original codon, in Tables 10a and 11a, it will be noted that the nonrandom process is less efficient: it takes more base replacements to achieve the same observable result as for fewer replacements in the random case: thus to achieve an average of 2.18 observable replacements per codon requires only six total replacements during random

Table 9a. α Hemoglobin Genes or mRNA (p_1:p_2:p_3::0.12:0.12:0.76)

Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.187973	0.205627	0.400774	0.393600	0.
3	1.354905	0.098147	0.514465	0.321724	0.065664
4	1.493365	0.100849	0.390110	0.423866	0.085174
5	1.608808	0.075443	0.369211	0.426441	0.128905
6	1.705105	0.066041	0.323779	0.449214	0.160966
7	1.785485	0.055428	0.297576	0.453078	0.193917
8	1.852622	0.048555	0.272323	0.457068	0.222054
9	1.908735	0.042735	0.253126	0.456808	0.247331
10	1.955664	0.038343	0.236793	0.455721	0.269143
11	1.994938	0.034790	0.223528	0.453637	0.288045
12	2.027825	0.031972	0.212467	0.451324	0.304237
13	2.055383	0.029690	0.203317	0.448913	0.318080
14	2.078490	0.027840	0.195692	0.446607	0.329862
15	2.097876	0.026327	0.189340	0.444464	0.339870
16	2.114150	0.025086	0.184033	0.442528	0.348354
17	2.127821	0.024061	0.179594	0.440808	0.355537
18	2.139311	0.023214	0.175875	0.439298	0.361614
19	2.148975	0.022509	0.172757	0.437983	0.366750
20	2.157107	0.021923	0.170139	0.436847	0.371091
∞	2.200832	0.018856	0.156145	0.430311	0.394688

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively <θ> is the average number of sites/codon which differ from the homologous sites in the original codon.

All values in this Table were calculated from Eq. 4 using the primary data in legend to Table 1.

divergence but eight total replacements for the nonrandom situation existing in α hemoglobin. Thus nonrandom replacement is not only less efficient at a given nucleotide site within the gene as shown in the preceding section, but also within a codon. It should also be noted that for both Tables 10a and 11a the observable effects are near their asymptotic values (the codon is "saturated" with hits) after about ten fixed mutations.

Table 10a. α Hemoglobin Genes or mRNA ($p_1 = p_2 = p_3 = 1/3$)
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.540398	0.126269	0.207064	0.666667	0.
3	1.834903	0.023670	0.339979	0.414129	0.222222
4	1.996700	0.042647	0.194092	0.487175	0.276086
5	2.086268	0.023065	0.204649	0.435238	0.337048
6	2.136208	0.025176	0.171624	0.445016	0.358184
7	2.164240	0.020668	0.170195	0.433364	0.375772
8	2.180073	0.020681	0.161719	0.434444	0.383155
9	2.189068	0.019523	0.160403	0.431556	0.388518
10	2.194206	0.019401	0.158034	0.431522	0.391042
11	2.197155	0.019079	0.157430	0.430746	0.392744
12	2.198855	0.019013	0.156726	0.430654	0.393607
13	2.199840	0.018918	0.156487	0.430432	0.394163
14	2.200413	0.018890	0.156268	0.430383	0.394460
15	2.200747	0.018860	0.156178	0.430316	0.394646
16	2.200942	0.018849	0.156108	0.430295	0.394748
17	2.201057	0.018840	0.156075	0.430273	0.394812
18	2.201125	0.018836	0.156051	0.430265	0.394848
19	2.201165	0.018833	0.156040	0.430258	0.394870
20	2.201188	0.018831	0.156031	0.430255	0.394882
∞	2.201223	0.018829	0.156021	0.430250	0.394901

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. $<\theta>$ is the average number of sites/codon which differ from the homologous sites in the original codon.

All values in this Table were calculated from Eq. 4 using the primary data in legend to Table 2.

If the distribution of fixed mutations over each codon is not (probablistically) even, but concentrated at the third position within the codon, then (Table 9a) the observable quantitative effects can be significant in magnitude: for n = 4, the expected proportion of codons with one observable base replacement rises to 39.01% because most – 3.08 (= 4 x 0.76) on the average – of these hits are superimposed on one another at the third position within the codon (recalling that $p_1 = p_2 = 0.12$, $p_3 = 0.76$). Moreover, there continue to be observable effects out to about 20 base replacements per codon. It is important to note that for such large numbers of replacements the asymptotic values are much the same for both random and nonrandom molecular divergence as can be seen from the last row in Tables 9a–16a.

Table 11a.
Both transitions and transversions permitted: random probability $P_n(\theta)$, corrected for multiple hits at the same base site and for revertants, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.555555	0.111111	0.222222	0.666667	0.
3	1.864197	0.024691	0.308643	0.444444	0.222222
4	2.035665	0.034294	0.192044	0.477366	0.296296
5	2.130925	0.021338	0.183051	0.438958	0.356653
6	2.183847	0.020339	0.159562	0.436011	0.384088
7	2.213248	0.017729	0.152689	0.428187	0.401395
8	2.229582	0.016965	0.146813	0.425896	0.410326
9	2.238657	0.016313	0.144234	0.423938	0.415516
10	2.243698	0.016026	0.142573	0.423078	0.418323
11	2.246499	0.015841	0.141726	0.422524	0.419908
12	2.248055	0.015747	0.141230	0.422242	0.420780
13	2.248919	0.015692	0.140963	0.422077	0.421267
14	2.249400	0.015663	0.140812	0.421988	0.421537
15	2.249666	0.015646	0.140729	0.421937	0.421687
16	2.249815	0.015637	0.140683	0.421910	0.421771
17	2.249897	0.015631	0.140657	0.421894	0.421817
18	2.249943	0.015629	0.140643	0.431886	0.421843
19	2.249968	0.015627	0.140635	0.421881	0.421857
20	2.249982	0.015626	0.140631	0.421878	0.421865
∞	2.250000	0.015625	0.140625	0.421875	0.421875

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. $<\theta>$ is the average number of sites/codon which differ from the homologous sites in the original codon.

All values in the Table were calculated from Eq. 4 using the primary data in the legend to Table 3.

Again by comparing the values for $<\theta>$ in Tables 9a and 10a it can be seen that genetic change is most efficient (the greatest amount of observable change per base replacement) for purely random mutation: the $<\theta>$ values in Table 11a are the largest for a given number of base replacements within the codon. Thus nonrandomness in the distribution of fixed mutations within the codon also reduces the efficiency of molecular change.

In Table 12a, $P_n(\theta)$ is given for β hemoglobin, the base composition being based on the observed base sequence of mRNA sequences of rabbit and human β hemoglobin (Kafatos et al. 1977) at the 119 codon loci corresponding to the amino acid residues observed to have varied in the β hemoglobin from 59 species. In Table 13a, also for

Table 12a. β Hemoglobin Genes or mRNA ($p_1:p_2:p_3::0.12:0.12:0.76$)
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.123506	0.270094	0.336306	0.393600	0.
3	1.306121	0.097133	0.565276	0.271926	0.065664
4	1.438413	0.134361	0.365276	0.427952	0.072411
5	1.551270	0.083771	0.404295	0.388828	0.123106
6	1.644172	0.085338	0.328551	0.442711	0.143400
7	1.721543	0.065807	0.325064	0.430910	0.178220
8	1.785934	0.061932	0.289769	0.448732	0.199567
9	1.839636	0.052753	0.278746	0.444613	0.223888
10	1.884486	0.048953	0.259348	0.449960	0.241740
11	1.922000	0.043975	0.249061	0.447952	0.259012
12	1.953422	0.041086	0.237135	0.449050	0.272729
13	1.979780	0.038104	0.229102	0.447705	0.285089
14	2.001919	0.036041	0.221257	0.447444	0.295258
15	2.020540	0.034132	0.215327	0.446409	0.304132
16	2.036222	0.032683	0.209963	0.445801	0.311552
17	2.049446	0.031409	0.205671	0.444987	0.317934
18	2.060610	0.030391	0.201922	0.444373	0.323314
19	2.070046	0.029516	0.198834	0.443738	0.327912
20	2.078031	0.028798	0.196178	0.443218	0.331806
∞	2.123714	0.024788	0.181146	0.439631	0.354435

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. <θ> is the average number of sites/codon which differ from the homologous sites in the original codon.
The average nucleotide composition for the first position was <A> = 0.2238, <C> = 0.2069, <G> = 0.4506 and <U> = 0.1187. In the second position it was <A> = 0.3235, <C> = 0.2311, <G> = 0.1471 and <U> = 0.2983. In the third position it was <A> = 0.0683, <C> = 0.2784, <G> = 0.3581 and <U> = 0.2952.
All values in this Table were calculated from Eq. 4 using the data in legend to Table 4.

β hemoglobin, the base composition at the first two positions was inferred from the amino acid composition in these 59 species at the 119 varied loci, the rationale for this being that 59 species might give a more representative composition than only two (human and rabbit). For Table 13a at the third position within the codon we continued to use the base composition from the human/rabbit mRNA sequences, because the degeneracy of the genetic code is so extensive at the third position that the base composition there cannot be inferred meaningfully from amino acid sequence data. As the tabular values in these two tables do not vary much from each other, Table 13a

Table 13a. β Hemoglobin Genes of mRNA (p_1:p_2:p_3::0.12:0.12:0.76)
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.123744	0.269856	0.336544	0.393600	0.
3	1.306741	0.097142	0.564639	0.272555	0.065664
4	1.439493	0.133983	0.365213	0.428133	0.072671
5	1.552842	0.083563	0.403448	0.389574	0.123415
6	1.646240	0.084906	0.327999	0.443044	0.144051
7	1.724089	0.065453	0.324027	0.431496	0.179023
8	1.788928	0.061471	0.288814	0.449030	0.200685
9	1.843044	0.052325	0.277509	0.444963	0.225203
10	1.888269	0.048476	0.258094	0.450115	0.243315
11	1.926119	0.043510	0.247640	0.448070	0.260780
12	1.957840	0.040600	0.235664	0.449033	0.274703
13	1.984460	0.037622	0.227524	0.447626	0.287228
14	2.006829	0.035550	0.219626	0.447268	0.297556
15	2.025650	0.033642	0.213625	0.446173	0.306560
16	2.041505	0.032189	0.208216	0.445495	0.314100
17	2.054878	0.030914	0.203874	0.444629	0.320582
18	2.066170	0.029895	0.200089	0.443966	0.326050
19	2.075716	0.029020	0.196967	0.443291	0.330722
20	2.083794	0.028301	0.194284	0.442736	0.334679
∞	2.129937	0.024291	0.179115	0.438962	0.357633

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. $\langle\theta\rangle$ is the average number of sites/codon which differ from the homologous sites in the original codon.

All values in the Table were calculated from Eq. 4 using the primary data in the legend to Table 5.

is probably to be preferred because of the much larger data base from which it is derived at the first two coding positions.

$P_n(\theta)$ is tabulated for myoglobin, cytochrome *c* and the parvalbumin group genes in Tables 14a, 15a, and 16a respectively.

Tables 9a, 12a or 13a, 14a, 15a, and 16a are the nonrandom analogues to Table 1 for random REH theory in Holmquist et al. (1972). The latter Table is reproduced in the present paper as Table 11a for convenience. The effect of the various nonrandomicities is perceptible. For example, for α hemoglobin (Table 9a), for two hits per codon, n = 2, the expected proportions of codons exhibiting no or a single base replacement is about doubled in the nonrandom case: in the random case $P_2(0) = 0.1111$

Table 14a. Myoglobin Genes of mRNA (p_1:p_2:p_3::0.12:0.12:0.76)
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.181518	0.212082	0.394318	0.393600	0.
3	1.345178	0.097512	0.525461	0.311362	0.065664
4	1.477943	0.106178	0.391016	0.421490	0.081316
5	1.587535	0.078208	0.380578	0.416687	0.124528
6	1.678062	0.070913	0.332532	0.444135	0.152420
7	1.753003	0.059719	0.310672	0.446496	0.183113
8	1.815166	0.053505	0.285662	0.452995	0.207838
9	1.866834	0.047596	0.268462	0.453455	0.230487
10	1.909863	0.043402	0.252912	0.454107	0.249579
11	1.945769	0.039841	0.240747	0.453214	0.266198
12	1.975790	0.037073	0.230375	0.452240	0.280311
13	2.000936	0.034781	0.221898	0.450925	0.292396
14	2.022040	0.032929	0.214765	0.449642	0.302663
15	2.039782	0.031397	0.208832	0.448364	0.311407
16	2.054724	0.030134	0.203843	0.447187	0.318836
17	2.067330	0.029084	0.199657	0.446106	0.325154
18	2.077982	0.028208	0.196127	0.445139	0.330526
19	2.086998	0.027475	0.193149	0.444281	0.335096
20	2.094639	0.026859	0.190629	0.443526	0.338986
∞	2.139358	0.023352	0.175976	0.438633	0.362039

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. <θ> is the average number of sites/codon which differ from the homologous sites in the original codon.

All values in the Table were calculated from Eq. 4 using the primary data in the legend to Table 6.

and $P_2(1) = 0.2222$; for the nonrandom case $P_2(0) = 0.2056$ and $P_2(1) = 0.4008$. On the other hand in the nonrandom case many fewer codons have observable replacements at two positions within the codon: in the nonrandom case $P_2(2) = 0.3936$, while for random divergence $P_2(2) = 0.6667$. Although each gene family is nonrandom in a different way – different base compositions, different nucleotide transition probabilities (see legends to Tables 1–6) – it is rather unexpected that the observable effects of this nonrandom divergence are so similar for four of the five families. The tabular values for the genes for α hemoglobin, myoglobin, cytochrome *c* and the parvalbumin group are all very similar, and even those for the β hemoglobin gene are not all that different.

Table 15a. Cytochrome *C* Gene or mRNA (p_1:p_2:p_3::0.12:0.12:0.76)
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.184749	0.208851	0.397549	0.393600	0.
3	1.346810	0.097938	0.522978	0.313420	0.065664
4	1.479395	0.104702	0.392947	0.420607	0.081745
5	1.588659	0.078137	0.379429	0.418072	0.124362
6	1.678862	0.070294	0.333203	0.443850	0.152653
7	1.753484	0.059572	0.310348	0.447105	0.182975
8	1.815344	0.053242	0.285953	0.453023	0.207781
9	1.866730	0.047499	0.268510	0.453754	0.230237
10	1.909504	0.043301	0.253177	0.454240	0.249283
11	1.945182	0.039807	0.240985	0.453428	0.265781
12	1.975002	0.037055	0.230707	0.452417	0.279820
13	1.999975	0.034795	0.222254	0.451133	0.291819
14	2.020929	0.032959	0.215172	0.449849	0.302020
15	2.038545	0.031444	0.209267	0.448589	0.310701
16	2.053383	0.030194	0.204309	0.447419	0.318079
17	2.065903	0.029153	0.200144	0.446350	0.324353
18	2.076485	0.028285	0.196633	0.445392	0.329689
19	2.085445	0.027558	0.193669	0.444542	0.334231
20	2.093044	0.026947	0.191161	0.443794	0.338098
∞	2.137884	0.023435	0.176451	0.438908	0.361206

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. $\langle\theta\rangle$ is the average number of sites/codon which differ from the homologous sites in the original codon.

All values in the Table were calculated from Eq. 4 using the primary data in the legend to Table 7.

These results suggest that the net effect of nonrandomness in these five families can be represented by a single table derived from the average values for the five gene families. Table 17a is the result and is a useful summary of the data in Tables 9a–16a.

4.3. Expected Proportion of Variable Amino Acid Sites with no, one, two, or three Minimal Base Replacements After n *Fixed Hits per Codon During Nonrandom Molecular Divergence.* In two proteins, at a given amino acid residue, the type of replacement that in fact occurs in the genes is reflected as an amino acid replacement in the proteins. Thus the actual two-base codon replacement UGC ↔ AGA is a replacement of cysteine by arginine, Cys ↔ Arg, at the protein level. This cysteine to arginine re-

Table 16a. Parvalbumin group genes (p_1:p_2:p_3::0.12:0.12:0.76)
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.185789	0.207811	0.398589	0.393600	0.
3	1.347675	0.098014	0.521962	0.314361	0.065664
4	1.480520	0.104127	0.393243	0.420613	0.082017
5	1.590063	0.077931	0.378622	0.418900	0.124547
6	1.680570	0.069880	0.332794	0.444204	0.153123
7	1.755505	0.059248	0.309507	0.447738	0.183507
8	1.817675	0.052848	0.285148	0.453484	0.208520
9	1.869362	0.047117	0.267510	0.454267	0.231106
10	1.912421	0.042890	0.252117	0.454675	0.250318
11	1.948363	0.039389	0.239813	0.453844	0.266954
12	1.978423	0.036622	0.229465	0.452780	0.281133
13	2.003613	0.034353	0.220934	0.451459	0.293254
14	2.024762	0.032507	0.213792	0.450132	0.303568
15	2.042549	0.030985	0.207832	0.448833	0.312350
16	2.057535	0.029728	0.202828	0.447626	0.319818
17	2.070184	0.028682	0.198623	0.446523	0.326172
18	2.080877	0.027810	0.195079	0.445534	0.331577
19	2.089931	0.027080	0.192087	0.444655	0.336178
20	2.097608	0.026467	0.189555	0.443881	0.340097
∞	2.142699	0.022972	0.174792	0.438802	0.363434

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. $<\theta>$ is the average number of sites/codon which differ from the homologous sites in the original codon.

All values in the Table were calculated from Eq. 4 using the primary data in the legend to Table 8.

placement will be recorded as a minimal *one* base (not two-base) replacement because it can be most parsimoniously, but incorrectly, explained by the codon change UGC ↔ CGC. Equation 7 in Section 3.2 corrects for this type of misclassification, the correction factors by which Tables 9a–16a must be adjusted (through Equation 7) being given in Tables 1–8. Comparing Tables 1 and 3, it is clear that nonrandom divergence significantly affects the final classification of amino acid replacements as minimal 0- (silent), 1-, 2-, and 3-base types. In the nonrandom case, and excluding chain-terminating interchanges, for α hemoglobin (Table 1), 56% of actual 1-base codon replacements at the gene level are observed as silent at the amino acid level, wheras for random divergence (Table 3) only 25% of those amino acid replacements are misclassi-

Table 17a. Composite summary for α hemoglobin, β hemoglobin, myoglobin, cytochrome c and parvalbumin group gene families (p_1:p_2:p_3::0.12:0.12:0.76)
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$, corrected for multiple hits at the same base site, for revertants and parallelisms, that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	<θ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.185472	0.208128	0.398272	0.393600	0.
3	1.350664	0.097946	0.519107	0.317283	0.065664
4	1.486501	0.103051	0.390872	0.422603	0.083474
5	1.599209	0.076726	0.374187	0.422238	0.126849
6	1.692804	0.068158	0.327952	0.446819	0.157072
7	1.770629	0.057378	0.303497	0.450243	0.188882
8	1.835423	0.050761	0.278514	0.455265	0.215459
9	1.889435	0.044934	0.260154	0.455454	0.239458
10	1.934517	0.040618	0.244239	0.455152	0.259992
11	1.972192	0.037071	0.231459	0.453678	0.277792
12	2.003715	0.034271	0.220736	0.451999	0.292993
13	2.030123	0.031986	0.211894	0.450132	0.305988
14	2.052271	0.030133	0.204505	0.448320	0.317042
15	2.070870	0.028611	0.198348	0.446602	0.326439
16	2.086504	0.027359	0.193192	0.445035	0.334414
17	2.099663	0.026322	0.188872	0.443627	0.341179
18	2.110749	0.025461	0.185244	0.442381	0.346914
19	2.120099	0.024743	0.182192	0.441288	0.351777
20	2.127993	0.024143	0.179621	0.440337	0.355900
∞	2.172355	0.020862	0.165263	0.434533	0.379342

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U), C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine and guanosine, respectively. <θ> is the average number of sites/codon which differ from the homologous sites in the original codon.

The average nucleotide composition for the first position was <A>= 0.2833, <C> = 0.1720, <G> = 0.4075 and <U> = 0.1372. In the second position it was <A>= 0.3757, <C> = 0.2471, <G> = 0.1261 and <U> = 0.2511. In the third position it was <A> = 0.2210, <C> = 0.2479, <G> = 0.2798 and <U> = 0.2513. The conditional nucleotide transition probabilities for the first position were A → C, G, U = 0.0667, 0.8869, 0.0464; C →A, G, U = 0.2739, 0.5313, 0.1949: G → A, C, U = 0.4835, 0.2943, 0.2222; and U → A, C, G = 0.2854, 0.2418, 0.4728. In the second position they were A → C, G, U = 0.4082, 0.1754, 0.4163; C → A, G, U = 0.6394, 0.1228, 0.2378; G → A, C, U = 0.4320, 0.2831, 0.2849; and U → A, C, G = 0.6501, 0.2311, 0.1118. In the third position they were A → C, G, U = 0.3113, 0.3719, 0.3168; C → A, G, U = 0.2771, 0.3941, 0.3288; G → A, C, U = 0.2945, 0.3489, 0.3567; and U → A, C, G = 0.2782, 0.3243, 0.3976. The relative frequencies with which the first, second and third position within the codon fixed mutations were taken to be 0.12, 0.12 and 0.76, respectively

Table 18a. Effect of a nonrandom distribution of hits within a codon on an otherwise random evolution of a gene
Both transitions and transversions permitted: nonrandom probability $P_n(\theta)$ corrected for multiple hits at the same base site, for revertants, and parallelisms that a codon hit n times has exactly θ nucleotide base sites which differ from the homologous sites in the original codon

n	$<\theta>$	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	1.000000	0.	1.000000	0.	0.
2	1.191467	0.202133	0.404267	0.393600	0.
3	1.360946	0.098318	0.508082	0.327936	0.065664
4	1.503193	0.097830	0.388699	0.425919	0.087552
5	1.622682	0.073498	0.362272	0.432279	0.131951
6	1.723053	0.063049	0.317471	0.452859	0.166621
7	1.807364	0.052550	0.288980	0.457026	0.201444
8	1.878186	0.045357	0.263053	0.459638	0.231953
9	1.937676	0.039495	0.242662	0.458513	0.259329
10	1.987648	0.035002	0.225561	0.456224	0.283213
11	2.029624	0.031418	0.211521	0.453081	0.303980
12	2.064884	0.028570	0.199846	0.449714	0.321870
13	2.094503	0.026278	0.190153	0.446356	0.337212
14	2.119382	0.024424	0.182078	0.443188	0.350309
15	2.140281	0.022915	0.175344	0.440286	0.361455
16	2.157836	0.021679	0.169720	0.437686	0.370915
17	2.172582	0.020663	0.165018	0.435392	0.378927
18	2.184969	0.019825	0.161083	0.433389	0.385703
19	2.195374	0.019132	0.157788	0.431655	0.391425
20	2.204114	0.018556	0.155027	0.430164	0.396253
∞	2.250000	0.015625	0.140625	0.421875	0.421875

A nucleotide site is said to have been hit each time a mutagenic event has occurred at that site. A mutagenic event is defined as a one-step change of one nucleotide to a different nucleotide; i.e. C → U = (C → U), C → U = (C → U → C → U). C → U (at any one nucleotide site) and A → G (at another site) represent, respectively, one, three and two mutagenic events. C, U, A and G are abbreviations for cytidine, uridine, adenosine, and guanosine, respectively. $<\theta>$ is the average number of sites/codon which differ from the homologous sites in the original codon. The average nucleotide composition for all three positions within the codon was <A> = 0.2500, <C> = 0.2500, <G> = 0.2500 and <U> = 0.2500. All conditional nucleotide transition probabilities (A → C, G, U = for example) were 1/3. The relative frequencies with which the first, second, and third position within the codon fixed mutations were taken to be 0.12, 0.12, and 0.76 respectively

Table 9b. α Hemoglobin Protein (p_1:p_2:p_3::0.12:0.12:0.76)
Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	<δ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.437781	0.562219	0.437781	0.	0.
2	0.632980	0.433947	0.499126	0.066927	0.
3	0.727043	0.389962	0.494993	0.113086	0.001960
4	0.829319	0.323564	0.526095	0.147799	0.002542
5	0.908302	0.286509	0.522527	0.187117	0.003847
6	0.977304	0.251798	0.523905	0.219494	0.004804
7	1.034477	0.226544	0.518221	0.249448	0.005787
8	1.082839	0.205556	0.512676	0.275142	0.006626
9	1.123344	0.188989	0.506060	0.297571	0.007381
10	1.157395	0.175446	0.499744	0.316778	0.008032
11	1.185966	0.164455	0.493720	0.333229	0.008596
12	1.209959	0.155431	0.488258	0.347232	0.009079
13	1.230102	0.148012	0.483367	0.359129	0.009492
14	1.247020	0.141879	0.479065	0.369212	0.009844
15	1.261232	0.136798	0.475315	0.377745	0.010142
16	1.273177	0.132573	0.472072	0.384959	0.010395
17	1.283219	0.129054	0.469283	0.391053	0.010610
18	1.291666	0.126115	0.466894	0.396199	0.010791
19	1.298774	0.123658	0.464855	0.400542	0.010944
20	1.304758	0.121599	0.463119	0.404209	0.011074
∞	1.336975	0.110658	0.453487	0.424077	0.011778

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

fied in this way. In general, the greater the nonrandomicity, the greater the misclassification. Thus, not only is nonrandom molecular divergence less efficient in requiring a greater number of base replacements to accomplish a given gene change, it also leads to greater inferential errors as to the nature of that gene change given only the proteins coded for by those changes. It is thus very important when using amino acid sequence data instead of mRNA data to modify Tables 9a–16a by the correction factors in Tables 1–8, before beginning the analysis. The results of this correction are given in Tables 9b–16b. Again comparing Tables 9b, 12b, or 13b, 14b, 15b, and 16b the tabular values are seen to be quite similar for the five different protein families and are thus well summarized in the single Table 17b.

Considering Table 10b (nonrandom base compositions and base transition probabilities) with Table 11b (random base compositions and base transitions) it is clear that classification of amino acid differences by their minimum base type can lead to erroneous factual conclusions. For a given $n \geqslant 1$ in Table 10b, the average minimum base

Table 10b. α Hemoglobin Protein ($p_1 = p_2 = p_3 = 1/3$)
Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	$<\delta>$	P(0)	P(1)	P(2)	P(3)
0	0	1.000000	0.	0.	0.
1	0.748220	0.251780	0.748220	0.	0.
2	1.070651	0.182406	0.564538	0.253056	0.
3	1.255861	0.112171	0.526428	0.354769	0.006632
4	1.351907	0.094956	0.466420	0.430385	0.008239
5	1.407152	0.077834	0.447237	0.464870	0.010058
6	1.437023	0.071728	0.430209	0.487374	0.010689
7	1.454191	0.066824	0.423374	0.498588	0.011214
8	1.463707	0.064723	0.418281	0.505562	0.011434
9	1.469195	0.063226	0.415946	0.509233	0.011594
10	1.472292	0.062512	0.414353	0.511465	0.011669
11	1.474087	0.062037	0.413559	0.512684	0.011720
12	1.475113	0.061794	0.413044	0.513416	0.011746
13	1.475712	0.061639	0.412773	0.513826	0.011763
14	1.476058	0.061556	0.412602	0.514071	0.011771
15	1.476261	0.061504	0.412508	0.514211	0.011777
16	1.476379	0.061475	0.412451	0.514294	0.011780
17	1.476449	0.061457	0.412418	0.514342	0.011782
18	1.476490	0.061447	0.412398	0.514371	0.011783
19	1.476514	0.061441	0.412387	0.514388	0.011784
20	1.476528	0.061438	0.412380	0.514398	0.011784
∞	1.476549	0.061433	0.412370	0.514413	0.011785

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

difference $<\delta>$ is greater than for the random case (Table 11b). On this basis one might be tempted to conclude the nonrandom process is more efficient than the random process in terms of the number of base replacements required to effect a given structural change in the protein. But from our earlier discussion to Tables 10a and 11a we know this conclusion is false: the nonrandom process is in fact *less* efficient. For the sets of data considered in this paper, Table 10b seems to be an isolated exception. For Tables 9b, 12b, or 13b, 14b, 15b, and 16b the average minimum base difference in each case leads to the correct inference that the nonrandom process is less efficient. Comparing Tables 9b and 10b for α hemoglobin, the fact that the average minimum base difference is greater for Table 10b than for Tables 9b or 11b is more due to the assumption of an even distribution of fixations within the codons in Table 10b than to parameters (such as base composition) inherent to the α hemoglobins or their genes themselves. This is a general result: the distribution of base replacements within codons, rather than nonrandom base compositions, base transition probabilities, or unequal usage of degenerate codons, dominates the observable consequences of nonrandom molecular divergence.

Table 11b
Both transitions and transversions permitted: random probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	<δ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.745247	0.254753	0.745247	0.	0.
2	1.047413	0.179628	0.593332	0.227041	0.
3	1.226851	0.112958	0.558869	0.316537	0.011636
4	1.319826	0.094012	0.507665	0.382808	0.015515
5	1.373398	0.078543	0.488192	0.414590	0.018675
6	1.402485	0.071717	0.474192	0.433979	0.020112
7	1.418866	0.067349	0.467454	0.444179	0.021018
8	1.427893	0.065116	0.463361	0.450037	0.021486
9	1.432932	0.063811	0.461204	0.453228	0.021758
10	1.435723	0.063107	0.459967	0.455021	0.021905
11	1.437277	0.062709	0.459292	0.456011	0.021988
12	1.438139	0.062491	0.458913	0.456563	0.022033
13	1.438618	0.062368	0.458704	0.456869	0.022059
14	1.438884	0.062301	0.458587	0.457039	0.022073
15	1.439032	0.062263	0.458523	0.457134	0.022081
16	1.439114	0.062242	0.458487	0.457186	0.022085
17	1.439160	0.062230	0.458467	0.457215	0.022087
18	1.439185	0.062224	0.458456	0.457232	0.022089
19	1.439199	0.062220	0.458449	0.457241	0.022090
20	1.439207	0.062218	0.458446	0.457246	0.022090
∞	1.439217	0.062216	0.458442	0.457252	0.022091

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

Tables 9b, 12b or 13b, 14b, 15b and 16b are the nonrandom analogues to Table 8 for random REH theory in Holmquist et al. (1972), the latter Table being reproduced as Table 11b here for ease of reference. The effects of nonrandomicities on observable protein structure are appreciable. In α hemoglobin (Table 9b) for three hits per n = 3, the expected proportion of amino acid identities for the nonrandom case is about 3.5 times that of the random case (0.39 *vs.* 0.11), although the expected proportion of amino acid replacements of the minimal 1-base type is not much changed (0.49 *vs.* 0.56). Minimal 2-base type replacement are about 1/3 (0.11 *vs.* 0.32) as frequent as randomly expected. Amino acid replacements of the minimal 3-base type, though rare, appear to be particularly sensitive quantitative indicators of nonrandomness, only 1/6th as many being expected for nonrandom as for random gene divergence (0.002 *vs.* 0.12).

With slightly different numbers the above conclusions for α hemoglobin during nonrandom divergence hold for all the protein families examined in this paper.

Table 12b. β Hemoglobin Protein
Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	<δ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.457689	0.542311	0.457689	0.	0.
2	0.626470	0.452850	0.467830	0.079320	0.
3	0.716444	0.404013	0.478165	0.115185	0.002636
4	0.825714	0.332933	0.511327	0.152833	0.002907
5	0.897935	0.303518	0.499971	0.191569	0.004942
6	0.968523	0.264080	0.509073	0.221090	0.005757
7	1.022361	0.242681	0.499431	0.250733	0.007155
8	1.070274	0.219705	0.498328	0.273955	0.008012
9	1.108899	0.204568	0.490952	0.295491	0.008988
10	1.142123	0.190271	0.487039	0.312985	0.009705
11	1.169529	0.179731	0.481407	0.328464	0.010398
12	1.192809	0.170389	0.477362	0.341300	0.010949
13	1.212224	0.163061	0.473098	0.352396	0.011445
14	1.228647	0.156754	0.469698	0.361695	0.011853
15	1.242428	0.151638	0.466507	0.369646	0.012209
16	1.254076	0.147287	0.463858	0.376348	0.012507
17	1.263890	0.143690	0.461494	0.382053	0.012763
18	1.272191	0.140644	0.459500	0.386877	0.012979
19	1.279207	0.138099	0.457760	0.390977	0.013164
20	1.285150	0.135944	0.456283	0.394453	0.013320
∞	1.319197	0.123801	0.447430	0.414540	0.014229

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

4.4. Relative importance of nonrandom amino acid composition and nonrandom nucleotide transition probabilities relative to the distribution of fixed mutations within codons. The fact that the evolutionary behavior of α hemoglobin, β hemoglobin, myoglobin, cytochrome *c* and the parvalbumin group of proteins, and their corresponding genes, can be discribed with good accuracy by a single summary table (Table 17a or 17b), despite their different biological functions, amino acid compositions and base transition probabilities suggests that for most protein or gene families it is the distribution of fixed nucleotide replacements within the codon that most affects the observable consequences of nonrandom molecular divergence. This was pointed out specifically for α hemoglobin in the preceding section. This conjecture is further supported by the fact that the set of parameters $\{p_1 = 0.12; p_2 = 0.12; p_3 = 0\ 76\}$ that describes the distribution of fixed mutations within codons is the only set held constant in the various Tables for each of the five families.

Table 13b. β Hemoglobin Protein
Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	<δ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.449006	0.550994	0.449006	0.	0.
2	0.616896	0.456525	0.470054	0.073421	0.
3	0.707460	0.409170	0.477369	0.110291	0.003170
4	0.816050	0.336623	0.514213	0.145656	0.003508
5	0.889125	0.307196	0.502442	0.184404	0.005958
6	0.959816	0.267154	0.512831	0.213061	0.006954
7	1.014345	0.245508	0.503280	0.242569	0.008642
8	1.062628	0.222199	0.502662	0.265451	0.009688
9	1.101800	0.206833	0.495405	0.286890	0.010872
10	1.135422	0.192319	0.491686	0.304249	0.011746
11	1.163254	0.181603	0.486129	0.319679	0.012589
12	1.186877	0.172110	0.482164	0.332465	0.013261
13	1.206619	0.164654	0.477939	0.343542	0.013866
14	1.223315	0.158239	0.474572	0.352825	0.014364
15	1.237342	0.153029	0.471399	0.360773	0.014799
16	1.249198	0.148600	0.468764	0.367472	0.015163
17	1.259194	0.144937	0.466408	0.373179	0.015476
18	1.267651	0.141835	0.464419	0.378006	0.015740
19	1.274800	0.139241	0.462683	0.382110	0.015966
20	1.280857	0.137046	0.461208	0.385589	0.016157
∞	1.315509	0.124686	0.452383	0.405666	0.017265

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

To test this conjecture the pattern of evolutionary divergence of a gene and its protein product, each evolving randomly in every respect *except* for the distribution of fixed mutations within the codon, was explicitly calculated. The results are in Table 18a and Table 18b. These Tables approximate the numerical values in Table 17 rather well, proving the conjecture and forming the factual basis for the general result given in the last sentence of the second paragraph of Section 4.3.

5. Application

A single concrete example may serve to illustrate the above calculations. For this purpose we have chosen the comparison of rabbit α hemoglobin and rabbit β hemoglobin. As the mRNA nucleotide sequences for these proteins have been published (Heindell et al. 1978 Efstratiadis et al. 1977) it will be possible to estimate the various evolutionary parameters from both the protein sequence data and from the nucleic acid data and compare the results.

Table 14b. Myoglobin Protein
Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	<δ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.511710	0.488290	0.511710	0.	0.
2	0.686044	0.407184	0.499588	0.093228	0.
3	0.782905	0.356139	0.507163	0.134353	0.002345
4	0.880814	0.299878	0.522334	0.174884	0.002904
5	0.955728	0.266795	0.515130	0.213629	0.004447
6	1.020531	0.236228	0.512456	0.245873	0.005443
7	1.073453	0.214386	0.504314	0.274760	0.006539
8	1.117953	0.196011	0.497447	0.299119	0.007422
9	1.154880	0.181715	0.489922	0.320132	0.008231
10	1.185794	0.169939	0.483241	0.337907	0.008913
11	1.211609	0.160438	0.477021	0.353034	0.009506
12	1.233246	0.152604	0.471556	0.365829	0.010010
13	1.251387	0.146168	0.466718	0.376671	0.010442
14	1.266632	0.140829	0.462519	0.385844	0.010809
15	1.279459	0.132691	0.458874	0.393611	0.011121
16	1.290270	0.032691	0.455735	0.400189	0.011386
17	1.299396	0.129591	0.453033	0.405764	0.011612
18	1.307112	0.126988	0.450716	0.410492	0.011804
19	1.313645	0.124796	0.448730	0.414507	0.011967
20	1.319184	0.122946	0.447029	0.417919	0.012106
∞	1.351633	0.112261	0.436774	0.438036	0.012929

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

Table 15b. Cytochrome *C* Protein

Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	<δ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.506073	0.493927	0.506073	0.	0.
2	0.682720	0.407521	0.502239	0.090241	0.
3	0.778740	0.358156	0.507795	0.131203	0.002847
4	0.876060	0.301338	0.524807	0.170311	0.003544
5	0.950904	0.268125	0.518238	0.208246	0.005391
6	1.015331	0.237629	0.516029	0.239725	0.006618
7	1.068072	0.215667	0.508525	0.267875	0.007932
8	1.112318	0.197348	0.501994	0.291651	0.009008
9	1.149060	0.183016	0.494890	0.312113	0.009981
10	1.179785	0.171266	0.488490	0.329437	0.010807
11	1.205444	0.161762	0.482555	0.344161	0.011522
12	1.226937	0.153942	0.477309	0.356618	0.012131
13	1.244958	0.147511	0.472671	0.367167	0.012651
14	1.260099	0.142180	0.468634	0.376093	0.013093
15	1.272838	0.137749	0.465132	0.383649	0.013469
16	1.283577	0.134050	0.462111	0.390049	0.013789
17	1.292644	0.130953	0.459512	0.395474	0.014061
18	1.300312	0.128351	0.457279	0.400078	0.014293
19	1.306807	0.126159	0.455364	0.403987	0.014489
20	1.312317	0.124308	0.453724	0.407311	0.014657
∞	1.344866	0.113526	0.443741	0.427075	0.015659

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

Table 16b. Parvalbumin group Proteins
Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	$<\delta>$	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.491371	0.508629	0.491371	0.	0.
2	0.672486	0.411942	0.503630	0.084428	0.
3	0.769288	0.364645	0.504721	0.127335	0.003299
4	0.867613	0.305673	0.525162	0.165045	0.004121
5	0.943938	0.272053	0.518213	0.203476	0.006258
6	1.009565	0.240795	0.516538	0.234973	0.007693
7	1.063545	0.218345	0.508984	0.263450	0.009220
8	1.108866	0.199588	0.502434	0.287501	0.010477
9	1.146598	0.184899	0.495216	0.308274	0.011612
10	1.178189	0.172853	0.488682	0.325888	0.012577
11	1.204613	0.163098	0.482603	0.340886	0.013413
12	1.226772	0.155071	0.477211	0.353593	0.014125
13	1.245372	0.148464	0.472435	0.364368	0.014734
14	1.261012	0.142985	0.468270	0.373493	0.015252
15	1.274182	0.138430	0.464650	0.381226	0.015693
16	1.285290	0.134627	0.461524	0.387780	0.016069
17	1.294673	0.131442	0.458831	0.393339	0.016388
18	1.302610	0.128767	0.456516	0.398058	0.016659
19	1.309335	0.126513	0.454529	0.402067	0.016891
20	1.315040	0.124611	0.452825	0.405476	0.017088
∞	1.348597	0.113600	0.442463	0.425677	0.018260

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

Table 17b. Composite summary for α hemoglobin, β hemoglobin, myoglobin, cytochrome *c* and parvalbumin group protein families
Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, of a codon hit n times is δ

n	<δ>	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.479698	0.520302	0.479698	0.	0.
2	0.663904	0.417649	0.500798	0.081553	0.
3	0.760756	0.369953	0.502122	0.125141	0.002784
4	0.861184	0.308969	0.524418	0.163075	0.003539
5	0.938991	0.274001	0.518384	0.202237	0.005378
6	1.006440	0.241548	0.517123	0.234670	0.006659
7	1.062076	0.218095	0.509742	0.264156	0.008007
8	1.108974	0.198533	0.503094	0.289239	0.009134
9	1.148114	0.183177	0.495684	0.310988	0.010152
10	1.180947	0.170597	0.488880	0.329500	0.011022
11	1.208443	0.160409	0.482517	0.345298	0.011777
12	1.231509	0.152034	0.476843	0.358701	0.012421
13	1.250864	0.145150	0.471809	0.370070	0.012972
14	1.267123	0.139452	0.467413	0.379694	0.013441
15	1.280792	0.134725	0.463597	0.387839	0.013839
16	1.292294	0.130789	0.460306	0.394728	0.014177
17	1.301981	0.127503	0.457477	0.400556	0.014464
18	1.310149	0.124752	0.455055	0.405486	0.014707
19	1.317041	0.122444	0.452984	0.409659	0.014913
20	1.322863	0.120504	0.451217	0.413191	0.015088
∞	1.355618	0.109741	0.440982	0.433195	0.016082

This table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text

The experimental data for the proteins α- and β hemoglobin are given at the top of Table 19. There are 82 amino acid differences, 61 of the minimal 1-base type and 21 of the minimal 2 base type. If we use Table 17b, and proceed to calculate those parameters which minimize chi-square (as in Section 3.4), the results of the middle portion of Table 19 are obtained. The following points should be noted. As the distribution of replacements among codons becomes more asymmetric, that is as the negative binomial parameter *r* decreases from plus infinity (the Poisson distribution, with mean $\mu = 4.42$) to unity (the geometric distribution, with mean $p/q = 5.15$), the *fixation intensity*, that is the average number of replacements per varion $\langle n \rangle \equiv REH_2/T_2$, increases as does the estimated number T_2 of varions themselves (going from 120 residues to 139 residues). It is worth commenting that the mean number $\langle n \rangle$ of replacements per varion can be interpreted as the product of two factors: a complexity factor *r*, the number of sources of mutability (see Eq. 16), and a damage factor *p/q*, the aver-

Table 18b. Effect of a nonrandom distribution of hits within a codon on the otherwise random evolution of a protein

Both transitions and transversions permitted: nonrandom probability $P_n(\delta)$ that the minimum base difference, relative to original unhit codon, hit n times is δ

n	δ	P(0)	P(1)	P(2)	P(3)
0	0.	1.000000	0.	0.	0.
1	0.449704	0.550296	0.449704	0.	0.
2	0.632641	0.433191	0.500976	0.065833	0.
3	0.724959	0.385575	0.507329	0.103657	0.003438
4	0.823786	0.321698	0.537403	0.136315	0.004584
5	0.900897	0.283302	0.539408	0.170380	0.006909
6	0.968129	0.248914	0.542769	0.199593	0.008725
7	1.024176	0.223094	0.540184	0.226173	0.010548
8	1.071654	0.201924	0.536643	0.249287	0.012146
9	1.111579	0.185027	0.531946	0.269448	0.013579
10	1.145221	0.171256	0.527096	0.286818	0.014830
11	1.173527	0.160036	0.522318	0.301728	0.015917
12	1.197343	0.150827	0.517856	0.314463	0.016854
13	1.217374	0.143246	0.513792	0.325305	0.017657
14	1.234216	0.136980	0.510167	0.334510	0.018343
15	1.248376	0.131786	0.506978	0.342309	0.018927
16	1.260278	0.127472	0.504201	0.348905	0.019422
17	1.270281	0.123880	0.501802	0.354477	0.019842
18	1.278687	0.120884	0.499741	0.359179	0.020196
19	1.285751	0.118383	0.497979	0.363142	0.020496
20	1.291686	0.116293	0.496478	0.366481	0.020749
∞	1.322862	0.105454	0.488320	0.384136	0.022090

This Table excludes homologous codon pairs at least one member of which is a terminating codon and was calculated from Eq. 7 in text.

age number of hits per source. These two factors are inversely related: as r decreases, p/q increases. The more asymmetric the distribution of replacements among codons becomes, the higher the total number of replacements must be, increasing from 530 for the Poisson distribution to 716 for the geometric distribution. Because of the experimental absence of amino acid replacements of the minimal 3-base type, although the agreement appears best for $r \cong 25$, with $\chi^2 = 0.729$, values of χ^2 for other values of r do not differ greatly from this and cannot be excluded with confidence. For $r = 25$, the expected number of amino acid replacements of each type were calculated from Eq. 23 and 24 These are given at the bottom of Table 19 and are in sensible agreement with the experimental values at the top of the Table. It should also be noted that 96% of the χ^2-value is due to the absence of observable minimal 3-base type replacements.

Table 19. Evolutionary parameters for rabbit α hemoglobin vs. β hemoglobin

Number of Observed Amino Acid Replacements with 0-, 1-, 2-, and 3- Minimum Base Differences.

0-	1-	2-	3-
57	61	21	0

r	$\frac{p}{q}$	$\langle n \rangle = \frac{rp}{q}$	T_2	REH_2	χ^2
1 (Geometric)	5.15	5.15	139	716	1.015
2.05	2.23	4.56	132	602	0.771
5.00	0.92	4.58	124	568	0.747
25.00	0.18	4.57	120	548	0.729
∞ (Poisson)	0	4.42	120	530	0.733

Expected Amino Acid Replacements with 0-, 1-, 2-, and 3- Minimum Base Differences for r = 25

0-	1-	2-	3-
57	60.9	20.5	0.7

Table 20. Evolutionary parameters for rabbit α hemoglobin mRNA vs. β hemoglobin mRNA

Number of Codons with 0, 1, 2, and 3 Actual Base Differences in the mRNAs

0-	1-	2-	3-
38	42	44	15

r	$\frac{p}{q}$	$\langle n \rangle = \frac{rp}{q}$	T_2	REH_2	χ^2
1 (Geometric)	6.91	6.91	124	857	0.773
2.05	3.03	6.21	116	720	0.252
5.00	1.15	5.75	112	644	0.018
10.00	0.56	5.57	111	618	0.001
25.00	0.22	5.43	111	602	0.023
∞ (Poisson)	0.	5.38	110	591	0.056

Expected Number of Codons with 0, 1, 2, and 3 Actual Base Differences in the mRNAs (for r = 10.00).

0-	1-	2-	3-
38.1	41.9	44.1	14.9

Table 21. A comparison of evolutionary parameters estimated by REH theory for rabbit α hemoglobin vs. rabbit β hemoglobin and their mRNAs under two differing assumptions: (1) genetic events occur equiprobably (random model), and (2) genetic events occur with unequal probability (nonrandom model).

	Estimation made from amino acid sequence data						
	Total minimal base differences	Total amino acid differences	r	p	μ_2[c]	T_2	REH_2
Random model	103	82	∞[a]	0[a]	2.01	115	231
Nonrandom model	103	82	25[b]	0.15[b]	4.57	120	548
	Estimation made from mRNA sequence data						
	Actual Base differenes	Total codon differences	r	p	μ_2	T_2	REH_2
Random model	175	101	∞[a]	0[a]	2.52	115	290
Nonrandom	175	101	10[b]	0.36[b]	5.57	111	618

[a] The distribution of fixed mutations among codons is assumed to be Poisson in the random model.

[b] The negative binomial parameters r and p are not assumed but estimated from the best fit to the data.

[c] $\mu_2 = rp/q$, $q = 1 - p$.

It is thus to be anticipated that one might be able to delimit the allowable values of the evolutionary parameters as well as to increase their accuracy and reduce the chi-square value by considering the mRNAs for α- and β hemoglobin in which base replacements at all three positions within the codons are explicitly expressed. This is illustrated in Table 20.

There are a total of 175 nucleotide replacements between the mRNAs for α- and β hemoglobin. Of these, 42 codons have a single base replacement, 44 have two base replacements and 15 have three base replacements. Comparing Tables 19 and 20, it is obvious that the pattern of actual base replacements can be quite different from the pattern of minimal base differences. For determining the evolutionary parameters from mRNA data, we use Table 17a and proceed as before. The results are in the middle section of Table 20. Although the parameters in Table 20 are more accurate than those in Table 19 because actual rather than minimal base differences were used, the values in Table 19, derived solely from protein sequence data, are quite reasonable approximations to the more accurate values. Thus when gene or mRNA data is lacking, the information in the protein sequences seems sufficient for use provided the "b" rather than "a" tables are used and provided the sequences are distantly enough related that the types and number of amino acid substitutions observed represent a fair statistical sampling. For Table 20, the lowest χ^2 is 0.001, less than for Table 19, and delineates $r \cong 10$ and $p/q \cong 0.56$ as the distribution among codons that best describes the data. The estimated fixation intensity 5.57 base replacements per varion is somewhat higher than that (4.57) suggested by the protein data, and the estimated number of varions $T_2 = 111$ is somewhat less than the protein data esti-

mate (120). Overall the total number of replacements is more when estimated from the mRNA sequences than when estimated from the protein sequence data, 618 *vs* 548.

The bottom of Table 20 shows the expected distribution of replacements for $r = 10.0$ is in good agreement with the observed values at the table's top.

It is worthwhile to compare these values with those predicted by the original REH model (Jukes and Holmquist 1972) which assigns all the parameters their random values. An analysis of rabbit α- and β hemoglobin and of their corresponding mRNAs has been published (Holmquist 1980) for the case where genetic events are taken as equally probable as in the original REH model. The results from this random model are compared with the results of the present paper (Tables 19 and 20) in Table 21. We note the following. First, with respect to the raw data, minimal base differences are a poor approximation to the actual number of base differences observed between the mRNAs (103 *vs*. 175); and the number of amino acid differences do not accurately reflect total codon differences (82 *vs*. 101). Nevertheless, the evolutionary parameters β_2, T_2 and REH_2, if estimated from the equations of REH theory are reasonably concordant whether the primary data are the amino acid sequences or the mRNAs. For less distantly related sequences we should not expect the agreement to be so good: T_2 more generally would be severely underestimated from the amino acid sequence data because the degeneracy of the genetic code would not reveal most changes that had occurred at the third position within the codons. Secondly, the commonly assumed Poisson distribution of fixed mutations among codons gives a poorer fit to the data. This is most apparent from the mRNA sequence data where the negative binomial parameters are $\{r, p\} = \{10, 0.36\}$; these parameters would be $\{\infty, 0\}$ for a Poisson process. A consequence of this uneven distribution of fixed mutations among codons is a much increased fixation intensity (average number of fixed mutations per varion) 2.52 for an assumed Poisson process, and 5.57 for the best fit to the data within the limits of Eq. 14. Finally, using the mRNA data, and allowing for the nonrandomicities in the mRNA structures, and during the process of mutation and subsequent fixation shows that the rabbit α- and β hemoglobin mRNAs are separated by approximately 618 fixed mutations rather than the 231 estimated from the amino acid sequence data and a random model.

6. Discussion

The basic result of this paper, aside from the methodology, is that nonrandomness of any sort introduces an inefficiency in passing from a given gene or protein structure to another, say from an ancestral gene or protein to a contemporary one. This inefficiency manifests itself in that more base replacements are required to effect that passage than would be the case if evolutionary events of a given type were equiprobable. The fact that not all evolutionary events of a given type have an equal probability of occurrence is the result of natural selection for function. A greater number of base replacements to effect a given change in observable molecular structure is the molecular price paid for this selection.

The principle source of this inefficiency is the uneven distribution of fixed mutations *within* and *among* codons. The pattern of codon usage, amino acid composition, and the relative frequency with which a particular base is replaced by another also con-

tribute to this inefficiency but to a much lesser extent; this is fortunate because the evolutionarily important base transitions occur between experimentally inaccessible ancestral sequences or between an inaccessible ancestral sequence and a contemporary sequence. These transition frequencies thus cannot be directly determined and must be estimated by any of several procedures, each of which though appearing plausible in its own right is open to some debate. The amino acid or base compositions of the ancestral sequences are also, of course, not experimentally accessible. No serious error is made by assuming A:C:G:U:1:1:1:1 or by taking all conditional base transition probabilities to be 1/3, although that procedure is not recommended if better data from gene or mRNA sequence are available.

The present paper also throws some light on the question of whether or not codons that can mutate by one, two, or three base replacements to a chain terminator are selected against. By comparing the entries for "Term-Amino-Acid" in Table 3 to the corresponding entries in Tables 1, 4, 5, 6, 7, and 8, it is clear there is selection against each of these three types of base replacements.

The distribution of base replacements among codons was investigated with the aid of the two parameter (r, p) negative binomial distribution, in part because it had been found useful by Uzzell and Corbin in describing the replacements of the cytochromes c, but also because it contained two limiting forms: the one parameter (μ) Poisson distribution, which has often been used to describe the pattern of replacements in proteins and nucleic acids, and the geometric distribution which is also fully defined by a single parameter (p) but which does not seem to have been discussed in the literature before. The qualitative shapes of these two limiting distributions are quite different if the mean number of fixed mutations per codon is greater than one. For the Poisson distribution the probability of a codon sustaining n base replacements is largest near the mean value $\langle n \rangle$. In the geometric distribution on the other hand, the probability of a codon sustaining n base replacements is largest at $n = 0$, and decreases monotonically as n increases, the rate of decrease being slower the larger the mean value $\langle n \rangle$. The negative binomial distribution, for the parameter values of immediate interest is of intermediate character. The probability of a codon sustaining n base replacements is largest at some value n_{max} in the range $0 < n_{max} < \langle n \rangle$ but falls off more slowly on either side of n_{max} remaining appreciable in magnitude for quite large values of n. This asymmetry is responsible for the overall greater number of replacements necessary to achieve a given structural change. For the rabbit α-/β hemoglobin gene divergence the negative binomial parameters that gave the best fit to the mRNA sequence data were $r \cong 10.0$ and $p \cong 0.36$. These may be compared to the values

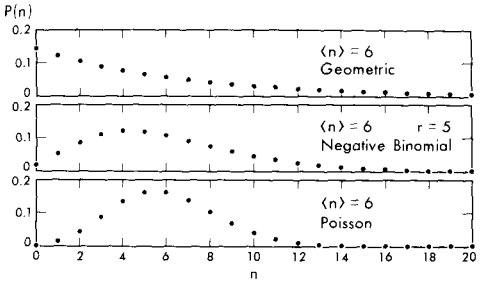

Fig. 2. Proportion P(n) of codons sustaining exactly n base replacements under three different distribution of replacements among codons: top (geometric), middle (negative binomial), botton (Poisson). The average number of replacements per codon was $\langle n \rangle$ = 6 in all three cases

found for the cytochromes *c* by Uzzell and Corbin (1971) $r \cong 2.05$ and $p \cong 0.67$. From these latter values the average fixation intensity (rp/q) was 4.51 base replacements per varion, which compares to the value of 5.57 base replacements per varion for the rabbit α-/β divergence.

Based on the pattern of base replacements within codons for the mRNAs or rabbit and human β hemoglobin, the relative frequency of occurrence p_1, p_2, p_3 of fixed replacements at the first, second, and third position within the codon was taken to be about 0.12, 0.12, and 0.76 respectively. There is nothing ultimately sacred about these numbers in the sense that one need not expect all genes to follow this same pattern. It is important to note that when estimating these parameters from mRNA or DNA sequence data it is necessary to consider closely, not distantly related species. For the latter, mutational saturation may have occurred: each position within the codon will have an observable difference between mRNAs of one base replacement at each of the three codon positions. It would thus falsely appear that $p_1 \cong p_2 \cong p_3$. This washing out of positional effects can be avoided by considering more closely related genes or mRNAs. Another manner of estimating these parameters is to let them freely vary, along with r, p, and T_2 during the minimization of chi-square. To accomplish this minimization in practice, we note that the total number N of observed base replacements between two mRNAs (or DNAs) can be decomposed in two manners:

$$N = N_1 + N_2 + N_3 \tag{25}$$

$$N = M_1 + 2M_2 + 3M_3 \quad , \tag{26}$$

where N_k is the number of observed base replacements at the k^{th} position within the codons (k = i, 2, and 3), and M_k is the number of codons having exactly k observed base replacements. Equating these two expressions,

$$N_1 + N_2 + N_3 = M_1 + 2M_2 + 3M_3 \tag{27}$$

Equation 27 contains five independent observable quantities. These should suffice to determine the five parameters that we wish to estimate: r, p, T_2 and p_1, p_2 ($p_3 = 1 - p_1 - p_2$), by requiring the expected N_k, as well as the expected M_k to agree with the observed values as closely as possible. In Section 5, only the expected and observed M_k were required to agree (see Table 20), the p_k being assumed ($p_1 = p_2 = 0.12$, $p_3 = 0.76$). By using the additional experimental information given by the N_k, the *a priori* assignment of values to the p_k as in Section 2.4 can be avoided. The expected values $\langle M_k \rangle$ required in the minimization procedure have been previously given in Equations 23 and 24:

$$\langle M_k \rangle \equiv n(k) \quad , \tag{28}$$

and the expected values of the N_k are given by

$$\langle N_k \rangle = T_2 \sum_{n=0}^{n=n} \max \Pr(n) \langle 1 - p(x_k) \rangle = T_2 \{1 - \sum_{n=0}^{n=n} \max \Pr(n) \langle p(x_k) \rangle\}$$

$$= T_2 \{1 - \sum_{n=0}^{n=n} \max \sum_{i,t=1}^{4} \Pr(n) B_{ik}\, {}^k\rho_{iit} [1 - p_k(1 - {}^k s_t)]^n\} \tag{29}$$

the expression for $\langle p(x_k) \rangle$ being given directly by Equation 6 and Pr(n) by Eq. 14. A computer program for implementing these refinements is currently being written. (See Note 1 added in proof.)

The proteins considered in this paper span a reasonably diverse range of biological function. The three globins are involved in oxygen transport and storage; cytochrome *c* is an electron transfer protein; and parvalbumin is a calcium binding protein. Chemically they are rather different: cytochrome *c* is a basic protein ($pI \cong 9$), the hemoglobins are neutral proteins ($pI \cong 6.5$) and the parvalbumins are acidic proteins ($pI \cong 4.5$). That despite these differences, the observable evolutionary behavior, with respect to those properties emphasized in this paper, is much the same is worth noting. It thus seems likely that the summary tables 17a and 17b should be applicable to other proteins as well as the five studied here. This result indicates that it is the mechanism (a constrained stochastic process) of molecular divergence, and not the details (the exact mole fraction of A, for example) that dominantes accurate estimates of genetic distance.

Lastly we note that though the present paper is couched in the language of the coding regions of genes, the methodology given is equally applicable to noncoding regions, such as introns, provided these regions are of sufficient length to be statistically informative. It is only necessary to divide these regions up into nonoverlapping triplets of bases so that the classification of base differences can be made as in Table 20, and Eqs. 25 and 26. Once this is done the analysis is the same as for a coding region.

One of the more important contributions of the present paper is the discipline imposed by the derivation of the explicit theoretical expressions characterizing nonrandom processes. Both the theoretical and numerical results show that the earlier understanding of molecular divergence considerably overestimated its efficiency.

Acknowledgments. This work was supported by NSF grant PCM76-18627 and NASA Contract NGR 05-003-460. We thank Mr. Thomas Conroy for assistance in the graphical drawing of Fig. 2, and G. William Moore for critical comments on the manuscript.

References

Barker WC, Ketcham LK, Dayhoff MO (1978) J Mol Evol 10:265–281

Bliss C, Fisher R (1953) Biometrics 53:176–200

Croft LR (1973) Handbook of protein sequences. Joynson-Bruvers, Oxford

Dayhoff MO (1972) Atlas of protein sequence and structure, vol 5. See also Supplement 1 (1973), Supplement 2 (1976) and Supplement 3 (1978)

Efstratiadis A, Kafatos FC, Maniatis T (1977) Cell 10:571–585

Feller W (1968) Elements of combinatorial analysis: binomial coefficients. In: An introduction to probability theory and its applications, vol 1 (2nd edition). John Wiley, New York, p 50

Feller W (1971a) Special densities, randomization. In: An introduction to probability theory and its applications, vol 2. John Wiley, New York, p 47, 57
Feller W (1971b) The exponential and the uniform densities. In: An introduction to probability theory and its applications, vol 2. John Wiley, New York, p 8, 11
Fitch W, Markowitz E (1970) Biochem Gen 4:579–593
Fitch W (1976) Molecular evolutionary clocks. In: Ayala F (ed), Molecular evolution. Sinauer Associates, Sunderland Massachusetts, p 160
Fitch W (1976) J Mol Evol 8:13–40
Goodman M, Moore GW (1977) J Mol Evol 10:4–47, Tables 7, 8, and 9
Heindell HC, Liu A, Paddock GV, Studnicka GM, Salser WA (1978) Cell 15:43–54
Holmquist R (1972) J Mol Evol 1:115–133
Holmquist R (1976a) Random and nonrandom processes in the molecular evolution of higher organisms. In: Goodman M, Tashian R, Tashian J (eds), Molecular anthropology. Plenum Press, New York, p 89
Holmquist R (1976b) J Mol Evol 8:337–349
Holmquist R (1978a) J Mol Evol 11:361–374
Holmquist R (1978b) J Mol Evol 12:17–24
Holmquist R (1978c) J Mol Evol 11:225–231
Holmquist R (1978d) J Mol Evol 11:349–360
Holmquist R (1979) J Mol Biol 135:939–958
Holmquist R (1980) J Mol Evol 15:149–159
Holmquist R, Cimino JB (1980) BioSystems 12:1–22
Holmquist R, Cantor C, Jukes TH (1972) J Mol Biol 64:145–161
Holmquist R, Jukes TH, Moise H, Goodman M, Moore GW (1976) J Mol Biol 105:39–74
Holmquist R, Moise H (1975) J Mol Evol 6:1–14
Iizuka M, Ishiii K, Matsuda H (1975) J Mol Evol 5:249–254
Jukes TH, Holmquist R (1972) J Mol Biol 64:163–179
Kafatos F, Efstratiadis A, Forget B, Weissman S (1977) Proc Nat Acad Sci USA 74: 5618–5622, Figure 1 and Table 2
Karon J (1979) J Mol Evol 12:197–218
King M-C, Wilson AC (1975) Science 188:107–116
Levine RD, Tribus M (eds) (1979) The maximum entropy formalism. MIT Press, Cambridge
Marotta CA, Wilson JT, Forget BG, Weissman SM (1977) J Biol Chem 252:5040–5053
Mood AM (1950) Introduction to the theory of statistics, chapter 8. Point estimation. Mc-Graw Hill, New York, p 147
Moore GW, Goodman M, Callahan C, Holmquist R, Moise H (1976) J Mol Biol 105: 15–37
Nichols BP, Yanofsky C (1979) Proc Nat Acad Sci USA 76:5244–1979
Ortega JM, Rheinboldt WC (1970) Iterative solution of nonlinear equations in several variables. Academic Press, New York
Peacock D, Boulter D (1975) J Mol Biol 95:513–527
Schwartz R, Dayhoff MO (1978) Science 199:395–403
Uzzell T, Corbin K (1971) Science 172:1089–1096

Received May 20, 1979; Revised February 25, 1980

Note Added in Proof

If *only* functionally equivalent (Yockey 1977, *J. Theor. Biology* 67:337–343) residues are permitted to occupy a particular amino acid locus in a set of homologous proteins, thus excluding certain codons at that locus, this will introduce a type of nonrandomness only partially allowed for by the methods of this paper. To the extent this type of restriction causes different loci to fix different numbers of mutations and to the extent this is reflected in the number of observed base replacements required to go from the gene triplet coding for one member of that functionally equivalent amino acid residue to the triplet coding for another the effect is taken care of through the parameters r and p of the negative binomial distribution (Eq. 14) and through our classification of observed codon changes as being of the 0-, 1-, 2- or 3- actual base difference type as in Table 20.

Michael Coates and Simon Stone, in the Departments of Zoology and Botany of the University of Adelaide have recently studies the effect of a limited set of functionally equivalent residues on estimates of the total mutations fixed (*J. Mol. Evol.*, in press) in isolation from some of the other nonrandom factors of the present paper. Their result is important: restricting the number of functionally equivalent residues at a locus significantly increases the estimates of total fixed mutations. As an additional type of *nonrandomness*, this is in agreement with the general principle laid down in the present paper that any type of nonrandomness in the evolutionary process requires a greater number of base replacements to effect the passage from one gene structure to another. Quantitatively the effect analyzed by Coates and Stone is large and to our knowledge they are the first to explicitly calculate the quantitative consequences of this aspect of Darwinian selection.

Appendix

Coefficients and Arguments for Calculation of the Probability of Back Mutation (Eq. 1)

In statistical terminology, the four arguments s_j in Eq. 1 of the text are the reciprocals of the roots of a probability generating function. These roots are denominators in a partial fraction expansion that yields $P_{BB'}^{(X)}$ (Holmquist 1976b). These arguments s_j are independent of the base occupying a locus and sum to zero (because there is no linear term in the probability generating function). Each row and column of coefficients ρ_{iij} sum to unity, for when there have been no replacements, X = 0 in Eq. 1, the probability of back mutation is formally unity as the base at that locus has not changed. Because of rounding errors these identities in sums may not be perfectly exhibited by the tables. As the number X of replacements at a given locus becomes large in Eq. 1, all terms but the first vanish, and because s_1 is unity, the probabilities for back mutation approach the values in the first column of each table. These asymptotic values are, as common sense dictates, simply the average gene base composition. If X = 1, then the probabilities for back mutation given by Eq. 1 are zero, again in agreement with common sense, for if a base undergoes a one-step replacement to a different base, it clearly cannot remain the same. Because of this physical requirement that the probability for back mutation must vanish for X = 1, some coefficients or arguments may be negative. In general the coefficients and arguments are complex numbers though for the five gene families considered here all are real. The probability for back mutation is, of course, always a real number between zero and unity irrespective of whether or not particular terms on the right-hand side of Eq. 1 are complex or real. An annotated listing of the Fortran IV computer program which executes the calculation of the coefficients and arguments, given the base transition probabilities, is available to serious investigators.

The interested reader can compare the back mutation probabilities given by Eq. 1 with those for random mutation (Holmquist 1972; Iizuka et al. 1975):

$$m_{P_{ii}^{(X)}} = \frac{1}{4}\left[1 + \frac{(-1)^X}{3^{X-1}}\right] . \qquad \text{(A1)}$$

The effect on the probability for back mutation of nonrandom nucleotide transition probabilities relative to random transitions is illustrated in Fig. 1 (see main text of paper) for myoglobin with adenosine originally at the second position within the codon. For other proteins, other bases, or a different position within the codon, the behavior will differ in detail from that in Fig. 1. But it is clear that nonrandom transitions can have, in cases of practical interest, a marked effect on what is *observed* after multiple replacements.

In these Tables the three entries for each row and column *i, j* are for the first nucleotide position within the codon (top), second nucleotide position within the codon (middle), and third nucleotide position within the codon (bottom).

Table A1. α Hemoglobin

	Coefficients			
i/j	1	2	3	4
	0.234900	0.208711	0.295273	0.261116
A	0.289200	0.050827	0.338488	0.321485
	0.226303	0.739844	0.033853	0.
	0.216898	0.143295	0.000089	0.639718
C	0.322901	0.045769	0.022939	0.608392
	0.264698	0.035710	0.199592	0.500000
	0.379800	0.611769	0.008341	0.000090
G	0.124001	0.875170	0.000977	-0.000148
	0.244301	0.188735	0.566964	0.
	0.168402	0.036225	0.696297	0.099077
U	0.263899	0.028234	0.637597	0.070271
	0.264698	0.035710	0.199592	0.500000

		Arguments			
		s_1	s_2	s_3	s_4
1st	Position within codon	1.000000	-0.618323	-0.181152	-0.200526
2nd	Position within codon	1.000000	-0.141380	-0.351775	-0.506845
3rd	Position within codon	1.000000	-0.290589	-0.334161	-0.375250

The conditional nucleotide transition probabilities from which the values in this table were calculated are given in the legend to Table 1.

Table A2. β Hemoglobin

	Coefficients			
i/j	1	2	3	4
	0.223802	0.519926	0.021301	0.234971
A	0.323499	0.059737	0.060419	0.556346
	0.068299	0.930317	0.001162	0.000222
	0.206900	0.455960	0.137367	0.199773
C	0.231099	0.012180	0.742895	0.013827
	0.278400	0.011252	0.568918	0.141431
	0.450600	-0.002200	-0.007830	0.559430
G	0.147102	0.842862	0.009713	0.000324
	0.358100	0.032069	0.002981	0.606850
	0.118698	0.026314	0.849161	0.005827
U	0.298301	0.085222	0.186974	0.429504
	0.295201	0.026362	0.426940	0.251497

		Arguments			
		s_1	s_2	s_3	s_4
1st	Position within codon	1.000000	-0.060068	-0.132380	-0.807552
2nd	Position within codon	1.000000	-0.170884	-0.298399	-0.530717
3rd	Position within codon	1.000000	-0.072848	-0.342589	-0.584564

The conditional nucleotide transition probabilities from which the values in this table were calculated are given in the legend to Table 4. Codon usage and the equilibrium amino acid composition are those reported for the 119 varied codon loci of the mRNAs for human and rabbit β hemoglobins (Kafatos et al. 1977; Holmquist and Cimino 1980).

Table A3. β Hemoglobin

	Coefficients			
i/j	1	2	3	4
	0.221499	0.539596	0.003874	0.235031
A	0.329699	0.064861	0.047250	0.558190
	0.068299	0.930317	0.001162	0.000222
	0.179399	0.328390	0.357125	0.135086
C	0.217399	0.002535	0.765932	0.014134
	0.278400	0.011252	0.568918	0.141431
	0.446300	-0.009137	-0.001657	0.564493
G	0.155102	0.818088	0.025670	0.001140
	0.358100	0.032069	0.002981	0.606850
	0.152802	0.141151	0.640658	0.065390
U	0.297800	0.114516	0.161148	0.426536
	0.295201	0.026362	0.426940	0.251497

	Arguments			
	s_1	s_2	s_3	s_4
1st Position within codon	1.000000	-0.064454	-0.143462	-0.792084
2nd Position within codon	1.000000	-0.180257	-0.273153	-0.546589
3rd Position within codon	1.000000	-0.072848	-0.342589	-0.584564

The conditional nucleotide transition probabilities from which the values in this table were calculated are given in the legend to Table 5. Codon usage was calculated from the average amino acid composition at the 119 varied codon loci of 59 β hemoglobin chains on the assumption that the three positions within the codon behaved independently.

Table A4. Myoglobin

	Coefficients			
i/j	1	2	3	4
	0.307002	0.163296	0.151682	0.378020
A	0.447699	-0.000429	-0.008679	0.561408
	0.285699	-0.586417	0.	0.127885
	0.209198	0.039380	0.695854	0.055568
C	0.218102	0.513635	0.046128	0.222135
	0.222400	0.007451	0.500000	0.270148
	0.391100	0.039303	0.001221	0.568375
G	0.122302	0.004957	0.864667	0.008704
	0.269501	0.398681	0.	0.331819
	0.092700	0.758021	0.151242	-0.001963
U	0.211896	0.481836	0.097884	0.208383
	0.222400	0.007451	0.500000	0.270148

	Arguments			
	s_1	s_2	s_3	s_4
1st Position within codon	1.000000	-0.075804	-0.241855	-0.682341
2nd Position within codon	1.000000	-0.066831	-0.133594	-0.799574
3rd Position within codon	1.000000	-0.420164	-0.272470	-0.307366

The conditional nucleotide transition probabilities from which the values in this table were calculated are given in the legend to Table 6.

Table A5. Cytochromes *c*

	Coefficients			
i/j	1	2	3	4
	0.368735	0.126352	0.000629	0.504284
A	0.432141	-0.009782	0.007169	0.570472
	0.270733	0.599369	0.	0.129898
	0.126588	0.411820	0.461583	0.000010
C	0.262617	0.364371	0.080306	0.292705
	0.233947	0.007832	0.500000	0.258222
	0.363064	0.140485	0.000775	0.495675
G	0.103418	0.482468	0.418395	-0.004282
	0.261374	0.384968	0.	0.353659
	0.141613	0.321344	0.537013	0.000031
U	0.201824	0.162942	0.494129	0.141105
	0.233947	0.007832	0.500000	0.258222

	Arguments			
	s_1	s_2	s_3	s_4
1st Position within codon	1.000000	-0.035968	-0.242141	-0.721891
2nd Position within codon	1.000000	-0.075989	-0.167296	-0.756715
3rd Position within codon	1.000000	-0.381531	-0.294720	-0.323749

The conditional nucleotide transition probabilities from which the values in this table were calculated are given in the legend to Table 7.

Table A6. Parvalbumin group

	Coefficients			
i/j	1	2	3	4
	0.283299	0.384536	0.000414	0.331751
A	0.382699	0.030770	0.000125	0.586405
	0.254000	0.268190	0.	0.477810
	0.114302	0.292416	0.565239	0.028043
C	0.207399	0.033658	0.682496	0.076446
	0.240002	0.027728	0.500000	0.232270
	0.454899	-0.004295	0.000060	0.549336
G	0.129802	0.697666	0.170294	0.002238
	0.265997	0.676353	0.	0.057650
	0.147499	0.327343	0.434287	0.090870
U	0.280100	0.237906	0.147084	0.334910
	0.240002	0.027728	0.500000	0.232270

	Arguments			
	s_1	s_2	s_3	s_4
1st Position within codon	1.000000	-0.022016	-0.149739	-0.828245
2nd Position within codon	1.000000	-0.129015	-0.225183	-0.645802
3rd Position within codon	1.000000	-0.365454	-0.308080	-0.326466

The conditional nucleotide transition probabilities from which the values in this table were calculated are given in the legend to Table 8.

Table A7. Composite summary for α hemoglobin, β hemoglobin, myoglobin, cytochrome *c* and parvalbumin group gene families

	Coefficients			
i/j	1	2	3	4
	0.283300	0.340909	0.015918	0.359873
A	0.375701	0.075852	0.000006	0.598441
	0.221000	0.017580	0.760385	0.001035
	0.172000	0.228345	0.554743	0.044012
C	0.247098	0.042496	0.518160	0.192246
	0.247902	0.119183	0.120650	0.512265
	0.407501	0.006661	0.000164	0.585674
G	0.126101	0.872933	0.000453	0.000513
	0.279797	0.693816	0.024595	0.001792
	0.137199	0.424085	0.429175	0.009541
U	0.251100	0.058719	0.481381	0.208800
	0.251301	0.169421	0.094370	0.484908

	Arguments			
	s_1	s_2	s_3	s_4
1st Position within codon	1.000000	-0.087472	-0.217801	-0.694727
2nd Position within codon	1.000000	-0.143970	-0.234453	-0.621577
3rd Position within codon	1.000000	-0.392465	-0.281124	-0.326411

The conditional nucleotide transition probabilities from which the values in this table were calculated are given in the legend to Table 17a.