J.Mol.Evol. 8,337-349 (1976)

Journal of Molecular Evolution

Solution to a Gene Divergence Problem under Arbitrary Stable Nucleotide Transition Probabilities

RICHARD HOLMQUIST

University of California, Space Sciences Laboratory, Berkeley, Ca. 94720, USA

Received August 16, 1976

Summary. A nucleic acid chain L nucleotides in length, with the specific base sequence $B_1B_2....B_L$, each B_i being A, G, C, or T, is defined by the L-dimensional vector $B = (B_1, B_2, ..., B_L)$, the k^{th} position in the chain being occupied by the base B_k. Let $p_{BB'}$ be the twelve given constant non-negative transition probabilities that in a specified position the base B is replaced by the base B' in a single step, and let $P_{BB'}^{(X)}$ be the probability that the position goes from base B to B' in X steps. An exact analytical expression for $P_{BB'}^{(X)}$ is derived. Assuming that each base mutates independently of the others, an exact expression is derived for the probability $P_{BB'}^{(X)}$ that the initial gene sequence B goes to a sequence $B' = (B_1', B_2', ..., B_L')$ after $X = (X_1, X_2, ..., X_L)$ base replacements, where X_k is the number of single-step base replacements in the k^{th} position. The resulting equations allow a more precise accounting for the effects of Darwinian natural selection in molecular evolution than does the idealized but biologically less accurate assumption that each of the four nucleotides is equally likely to mutate to and be fixed as one of the other three. Illustrative applications of the theory to some problems in biological evolution are given.

Key words: Molecular Evolution/Homologous Proteins and Nucleic Acids/ Stochastic Evolutionary Models/ Amino Acid and Nucleotide Substitutions/ Computer Simulations of Molecular Divergence/Nucleotide Transition Probabilities/Gene Divergence

INTRODUCTION

In calculating the pathways of gene divergence during evolution, the only case which has been treated exactly is the idealized situation where each of the four types (A, G, C, or T) of nucleotides comprising the gene is equally likely to mutate to and be fixed as any one of the other three (Holmquist, 1972). The requirements of protein function (for structural genes), and the effects of Darwinian natural selection (for all genes)

during this divergence cause the probabilities with which one base mutates to and is fixed as one of the other three to differ (Goodman et al., 1974) from these equiprobable values. I derive here the equations necessary to calculate the divergence when these probabilities are arbitrary but constant. The results differ significantly from the idealized simplification which assumes equal probabilities.

THEORY

A fixation is defined as an accepted (by natural selection) point mutation, that is, a one-step replacement of one nucleotide by another. For example, A → G is one fixation, A → G → C → A are three fixations. Let the original base at the k^{th} nucleotide position be B_k (B_k = A, G, C, or T). After X_k fixations at that position, let the probability that the base B'_k (B'_k = A, G, C, or T) be found at that position be $P_{B_kB'_k}^{(X_k)}$. If the initial, i.e. ancestral, sequence of bases in the gene is defined by the vector $B = (B_1, B_2, \ldots, B_L)$, the probability $P_{BB'}^{(X)}$ that the final sequence is the vector $B' = (B'_1, B'_2, \ldots, B'_L)$ is given by

$$(1) \qquad P_{BB'}^{(X)} = \prod_{k=1}^{L} P_{B_kB'_k}^{(X_k)}$$

provided that each position mutates independently of the others. The vector $X = (X_1, X_2, \ldots, X_L)$ is the vector which describes the number of single-step base replacements which have occurred at each position.

The problem thus reduces to calculating $P_{B_kB'_k}^{(X_k)}$ for k = 1 to L. As the calculation is similar at each position we drop the subscript k in what follows. An explicit expression for $P_{BB'}^{(X)}$ will be derived by transforming an obvious recurrence relation into the generating function $G(s|P_{BB'})$ for $P_{BB'}^{(X)}$.

The probability that a given nucleotide position is occupied by the nucleotide B'(B' = A, G, C, or T) after the X^{th} fixation is equal to the sum of the conditional probabilities that if that position is *not* B' after the $(X - 1)^{th}$ fixation, it will mutate to and be fixed as B' at the next step:

$$(2) \qquad P_{BB'}^{(X)} = \sum_{B''} P_{BB''}^{(X-1)} \, p_{B''B'},$$

$$(3) \qquad \sum_{B' \neq B''} p_{B''B'} = 1, \text{ for a given } B''.$$

Here $p_{B''B'}$ is the probability that if a base is B" it will mutate to and be fixed as B'. Since it is biologically impossible for a base to mutate to itself, the four $p_{B''B'}$ with B" = B' (p_{AA}, p_{GG}, p_{CC}, p_{TT}) are zero. Equations (2) and (3) are valid for nonconstant as well as constant nucleotide transition probabilities $p_{B''B'}$; in this paper we derive the solution for $P_{BB'}^{(X)}$ when the $p_{B''B'}$ are constant and known. In Markov chain theory Equations (2) are known as the Chapman-Kolmogorov equations.

By definition the generating function $G(s|P_{BB'})$ for $P_{BB'}^{(X)}$ is (Feller, 1968)

$$(4) \qquad G(s|P_{BB'}) = \sum_{X=0}^{\infty} s^X P_{BB'}^{(X)}.$$

Multiplying each side of the recurrence relation (2) by s^X and summing each side over X = 1 to ∞, we obtain

$$(5) \qquad G(s|P_{BB'}) - P_{BB'}^{(0)} = s \sum_{B''} G(s|P_{BB''}) p_{B''B'}.$$

Here $P_{BB'}^{(0)}$ is the probability that the original base is B'. By assumption, the original base is B, so that $P_{BB}^{(0)} = 1$, and if $B' \neq B$, $P_{BB'}^{(0)} = 0$, since no more than one base can occupy a given nucleotide position. The only restriction on the variable s is that $G(s|P_{BB'})$ must converge in the interval $|s| < s_o$ for some s_o; s has no significance other than as an intermediate in the calculations.

For each starting base B, Equations (5) give four different equations, one for each of the four possible values of B' (A, G, C, or T). Since the single-step transition probabilities $p_{BB'}$ are constant and known, these four equations are linear with respect to the four unknowns $G(s|P_{BB'})$, and can be solved directly for the latter by Cramer's rule (Keller & Doherty, 1961) as the ratio of two determinants:

$$(6) \qquad G(s|P_{BB'}) = N(s|P_{BB'})/D(s).$$

D is the 4 × 4 determinant of the coefficients of $G(s|P_{BB'})$ in Eq. (5) and N is the 4 × 4 determinant obtained from D by replacing the coefficients of $G(s|P_{BB'})$ by the $P_{BB'}^{(0)}$ with sign changed, that is by -1, 0, 0, 0. D is the same for all B' and B, but the numerator N depends on both B' and B. By direct expansion of these determinants:

$$(7) \qquad D(s) = [1 - s][1 + s + \alpha s^2 + (\alpha - \beta)s^3]$$

where

$$(8)\qquad \alpha = 1 - p_{GT}p_{TG} - p_{CT}p_{TC} - p_{AT}p_{TA} - p_{CG}p_{GC} - p_{AG}p_{GA} - p_{AC}p_{CA}.$$

$$(9)\qquad \beta = (p_{TG}p_{GC}p_{CT} + p_{GT}p_{CG}p_{TC}) + (p_{TG}p_{GA}p_{AT} + p_{GT}p_{AG}p_{TA}) + (p_{CA}p_{AT}p_{TC} + p_{AC}p_{TA}p_{CT}) + (p_{CA}p_{AG}p_{GC} + p_{AC}p_{GA}p_{CG}).$$

If B' = B (the original base), then

$$(10)\qquad N(s|P_{BB'}) = 1 - s^{2}\,{}^{\Sigma^{*}(3)}_{B'',B'''\neq B}\, p_{B''B'''}p_{B'''B''} - s^{3}\,{}^{\Sigma^{*}(2)}_{B'',B''',B''''\neq B}\, p_{B''B'''}p_{B'''B''''}p_{B''''B'}$$

The number in parenthesis to the right of the asterisk in each summation is a reminder for the total number of terms in that sum, and the asterisk is a reminder that no term is to be taken more than once in the sum and that each term of the sum is to be independent of the other terms in that sum. If B'≠B (i.e. B' is one of the three nucleotides not originally occupying that locus),

$$(11)\qquad N(s|P_{BB'}) = s\left\{ p_{BB'} + s\,{}^{\Sigma^{*}(2)}_{B''} p_{BB''}p_{B''B'} + s^{2}\left[{}^{\Sigma^{*}(2)}_{\substack{B''\neq B'\\ B'''\neq B}} p_{BB''}p_{B''B'''}p_{B'''B'} - {}^{\Sigma^{*}(1)}_{B'',B'''\neq B,B'} p_{BB'}p_{B''B'''}p_{B'''B''} \right] \right\}$$

Explicit numerical examples of the expansions in Eqs. (10) and (11) are given in the section on illustrative applications.

Eqs. (6) through (11) completely define the generating function $G(s|P_{BB'})$ as explicit functions of the twelve $p_{B'B''}$ for all four B' (= A, G, C, or T). In Eq. (6) if N and D have common factors, let those be divided out expressing $G(s|P_{BB'})$ in the form

$$(12)\qquad G(s|P_{BB'}) = N^{*}(s|P_{BB'})/D^{*}(s),$$

with

$$(13)\qquad D^{*}(s) \equiv (s - s_1)(s - s_2)\ldots(s - s_k),$$

the s_i being the zeros of $D^{*}(s)$, and the asterisks indicating

that $N^*(s)$ and $D^*(s)$ have no common zeros. An explicit expression for $P_{BB'}^{(X)}$ for all nonnegative integral X is then given by the usual (Feller, 1968) partial fraction expansion of Eq.(12):

$$(14) \qquad P_{BB'}^{(X)} = \frac{\rho_1(B'|B)}{s_1^{X+1}} + \frac{\rho_2(B'|B)}{s_2^{X+1}} + \ldots + \frac{\rho_k(B'|B)}{s_k^{X+1}},$$

with

$$(15) \qquad \rho_i(B'|B) = - N^*(s_i|P_{BB'})/D^{*'}(s_i).$$

$N^*(s_i|P_{BB'})$ are the numbers obtained by evaluating $N^*(s|P_{BB'})$ at $s = s_i$, and $D^{*'}(s_i)$ are the numbers obtained by evaluating the first derivative of $D^*(s)$ [from Eq. (13)] with respect to s at $s = s_i$. Equations (14) and (15) hold provided $D^*(s)$ has no multiple zeros. The trivial modifications required when there are multiple zeros of $D^*(s)$ are in standard texts (Feller, 1968; D'Azzo & Houpis, 1966).

Equation (14) completes the analytical solution to our problem. Four applications, of increasing complexity, utility, and realism follow the Discussion.

DISCUSSION

The experimentally observed patterns of amino acid sequences of homologous proteins or nucleic acids which have been isolated from contemporary organisms are a function of, among other things, the probabilities for the transition of one nucleotide to another during the time over which these macromolecules diverged from a common ancestor. If the twelve transition probabilities can be estimated, then the expected distribution of nucleic acid and/or amino acid sequences can be predicted and compared with experiment.

The simplest estimate is to assume that each base has been equally likely to mutate to any one of the other three. In this case each of the twelve transition probabilities is simply 1/3 (Example 1, below). This model, despite its simplicity, has led to improved estimates of the number of nucleotide point fixations which separate the genes which code for proteins (Jukes & Holmquist, 1972), and which are in accord with those made by the method of parsimony (Moore et al., 1976; Holmquist et al., 1976). This concordance does not imply the absence of Darwinian natural selection, but rather that such has preserved a significant degree of randomness with respect to these transition probabilities.

Almost always the transition matrix elements for the genes coding for a particular protein family, and even within that

family for a particular phylogenetic lineage, will differ somewhat from the equiprobable values of 1/3, the actual values being specific for that family and lineage. In Example 2, below, we consider such a case and use it as a step by step concrete numerical example to show how the theoretical equations are employed in practice.

Example 3 illustrates how information at the gene level can be extended to the protein sequence level.

Finally, in the fourth example, the theory is applied to real data taken from the globin family of genes. The results demonstrate that the corrections are significant in magnitude and cannot be neglected in realistic biological situations.

ILLUSTRATIVE APPLICATIONS

Example 1. Probability of back mutation -- Consider the case where any base is equally likely to mutate to and be fixed as any one of the other three bases. Hence all $p_{B'B''} = 1/3$. What is the probability that after a fixations, the base is the same as at the start? For concreteness let the original base be G.

From Eqs. (7) and (10),

$$N(s|P_{GG}) = 1 - 3(1/3)(1/3)s^2 - 2(1/3)(1/3)(1/3)s^3 = -(2/27)\cdot(s+3)^2(s-\tfrac{3}{2})$$

$$D(s) = \left[1 - s\right]\left[1 + s + (1 - 6\cdot\tfrac{1}{3}\cdot\tfrac{1}{3})s^2 + (1 - 6\cdot\tfrac{1}{3}\cdot\tfrac{1}{3} - 8\cdot\tfrac{1}{3}\cdot\tfrac{1}{3}\cdot\tfrac{1}{3})s^3\right] = -(1/27)(s+3)^3(s-1).$$

Here not only do D and N have common zeros (-3), but D has a root (-3) of multiplicity three. From Eq. (6),

$$G(s|P_{GG}) = N(s|P_{GG})\ /D(s) = (2s - 3)/(s + 3)(s - 1)$$

in which the numerator and denominator have no common zeros and the denominator has no repeated roots. Thus $N^*(s|P_{GG}) = (2s - 3)$ and $D^*(s) = (s + 3)(s - 1)$ so that $D^{*'}(s) = 2(s + 1)$. From Eq. (15), and taking $s_1 = 1$, $s_2 = -3$,

$$\rho_1(G|G) = -N^*(1|P_{GG})\ /D^{*'}(1) = -(2\cdot 1 - 3)/2(1 + 1) = 1/4$$

$$\rho_2(G|G) = -N^*(-3|P_{GG})/D^{*'}(-3) = -(-3\cdot 2 - 3)/2(-3 + 1) = -9/$$

Finally, from Eq. (14),

$$P_{GG}^{(a)} = (1/4)/1^{a+1} - (9/4)/(-3)^{a+1} = \frac{1}{4}\left[1 + \frac{(-1)^a}{3^{a-1}}\right],$$

which is correct (Holmquist, 1972).

Example 2. Let the probabilities for the transition of one nucleotide to another be given by the matrix

	T	G	C	A
T	0	0.24	0.29	0.47
G	0.22	0	0.45	0.33
C	0.15	0.44	0	0.41
A	0.23	0.39	0.38	0

The numerical values of the matrix elements were chosen to illustrate the effect of moderate deviations from the overly idealistic situation in Example 1 where all nondiagonal matrix elements were 1/3. Consider a specific nucleotide position and let the ancestral base there be G. Derive explicit expressions for the probabilities that after k fixations the base there is G, T, C, or A respectively.
From Eqs. (7), (8), and (9),

$$\alpha = 1 - (0.22)(0.24) - (0.15)(0.29) - (0.23)(0.47) - (0.44)(0.45) - (0.39)(0.33) - (0.38)(0.41) = 0.3131.$$

$$\beta = (0.24)(0.45)(0.15) + (0.22)(0.44)(0.29) + (0.24)(0.33)(0.23) + (0.22)(0.39)(0.47) + (0.41)(0.23)(0.29) + (0.38)(0.47)(0.15) + (0.41)(0.39)(0.45) + (0.38)(0.33)(0.44) = 0.284082.$$

$$D(s) = (1 - s)(1 + s + 0.3131s^2 + 0.029018s^3)$$

$$= -0.0290(s - 1)(s + 6.010)(s - 2.395\ \underline{/3.080})\cdot$$

$$(s - 2.395\ \underline{/-3.080}),$$

$\underline{/\theta}$ being $\exp(i\theta)$ with θ in radians. From Eq. (10),

$$N(s|P_{GG}) = 1 - (p_{CT}p_{TC}+p_{AT}p_{TA}+p_{AC}p_{CA})s^2 - (p_{TC}p_{CA}p_{AT}+p_{CT}p_{AC}p_{TA})s^3$$

$$= 1 - [(0.15)(0.29) + (0.23)(0.47) + (0.38)(0.41)]s^2$$

$$- [(0.29)(0.41)(0.23) + (0.15)(.38)(0.47)]s^3$$

$$= 1 - 0.3074s^2 - 0.05414s^3 = -0.05414(s - 1.594)\cdot$$

$$(s + 2.359)(s + 4.913).$$

As $D(s)$ and $N(s|P_{GG})$ have no common factors $N^*(s|P_{GG}) \equiv N(s|P_{GG})$, and $D^*(s) \equiv D(s)$. The four zeros s_1, s_2, s_3, and s_4 of $D^*(s)$ are 1, -6.010, and +2.395 $\underline{/\pm 3.080}$.

$$D^{*\prime}(s) = -0.116s(s + 2.390)(s + 4.953).$$

From Eq. (15),

$$\rho_1(G|G) = -(-0.0541)(1 - 1.594)(1 + 2.359)(1 + 4.913)/(-0.116)\cdot(1)(1 + 2.390)(1 + 4.953) = 0.2726$$

$$\rho_2(G|G) = -N^*(-6.010|P_{GG})/D^{*\prime}(-6.010) = -0.6174$$

$$\rho_3(G|G) = -N^*(+2.395\ \underline{/3.080}|P_{GG})/D^{*\prime}(+2.395\ \underline{/3.080} = 0.7810\underline{/-2.911}$$

$$\rho_4(G|G) = 0.7810\ \underline{/+2.911}$$

Finally, from Eq. (14),

$$P_{GG}^{(k)} = 0.2726 + \frac{0.1027}{6.010^k}(-1)^k + \frac{0.6523}{2.395^k}\cos(5.991 + 3.080k).$$

The cosine term arises from the sum of the last two terms in Eq. (14), which involve the two complex conjugate roots, $2.395\ \underline{/\pm 3.080}$, of $D^*(s) = 0$. $P_{GA}^{(k)}$, $P_{GT}^{(k)}$, and $P_{GC}^{(k)}$ are calculated similarly with $N(s|P_{GA})$ being, for example, from Eq. (11):

$$N(s|P_{GA}) = s\{p_{GA}+s(p_{GC}p_{CA}+p_{GT}p_{TA})+s^2(p_{GC}p_{CT}p_{TA}+p_{GT}p_{TC}p_{CA}-p_{GA}p_{CT}p_{TG}$$

$$= s\left[0.33+\left[(0.45)(0.41)+(0.22)(0.47)\right]s+\left[(0.45)(0.15)(0.47)\right.\right.$$
$$\left.\left.+(0.22)(0.29)(0.41)-(0.33)(0.15)(0.29)\right]s^2\right.$$

$$= s(0.33+0.2879s+0.0435s^2) = 0.0435s(s+1.475)(s+5.139).$$

The results are

$$P_{GA}^{(k)} = 0.2824 - \frac{0.0644}{6.010^k}(-1)^k + \frac{2.100}{2.395^k}\cos(1.675 + 3.080k)$$

$$P_{GT}^{(k)} = 0.1667 - \frac{0.1245}{6.010^k}(-1)^k + \frac{0.5827}{2.395^k}\cos(4.640 + 3.080k)$$

$$P_{GC}^{(k)} = 0.2783 + \frac{0.0861}{6.010^k}(-1)^k + \frac{1.369}{2.395^k}\cos(4.443 + 3.080k).$$

In these solutions, for $k = 0$, $P_{GG}^{(0)}$, $P_{GA}^{(0)}$, $P_{GT}^{(0)}$, and $P_{GC}^{(0)}$ reduce to 1, 0, 0, 0 as required by the fact the initial base was G; also, for all $k \geq 0$, $P_{GG}^{(k)} + P_{GA}^{(k)} + P_{GT}^{(k)} + P_{GC}^{(k)}$ is identically unity as required by the fact that the nucleotide position must be occupied by some one of the four bases G, A, T, or C. For $k = 1$ the individual probabilities reduce to the elements of the transition matrix itself.

More interestingly, the terms $\rho_1(B'|B)$ in Eq. (14) which arise from the zero $s = 1$ in the generating function, are the

asymptotic values for the $P_{BB'}^{(k)}$ as $k \to \infty$. This asymptotic base composition is, for a given matrix of transition probabilities, the *only* base composition about which over the long term the nucleic acid composition can remain stable, but it should be noted that even if the starting nucleic acid has this asymptotic composition, it may first drift away from that composition before returning to it. The actual base composition is stable in the statistical sense, not in the sense of being constant. This is because the actual composition will fluctuate around the asymptotic expected values as a multinomial random variable with probabilities $P_{BB'}^{(\infty)}$. It should also be noted that a given asymptotic base composition does not uniquely determine the twelve nucleotide transition probabilities $p_{BB'}$, even though the converse is true.

The damped cosine term in the final expressions should not surprise us. If the matrix of transition probabilities is such that purine $\leftrightarrow$ pyrimidine base interchanges are forbidden then if the initial base is G, successive replacements result in the sequences G $\to$ A $\to$ G $\to$ A ad infinitum, with probabilities $P_{GG}^{(k)}$ of 0, 1, 0, ... (1 if k even, 0 if k odd). Thus $P_{GG}^{(k)}$ and $P_{GA}^{(k)}$ in such a case will exhibit undamped periodic behavior. Most biological situations will not be so extreme, but can be considered an admixture of various allowed and disallowed base transitions in various proportions so that a retention of some portion of the periodic component is natural.

Typically, each variable codon in a protein has received somewhere between one and four base replacements (Jukes & Homquist, 1972) over the geological time periods involved in genetic divergence. If we take X = 2 in Eq. (14), then, for this example, the expected base composition after two base replacements at the site initially occupied by G is $P_{GG}^{(2)} = 0.380$, $P_{GA}^{(2)} = 0.288$, $P_{GT}^{(2)} = 0.143$, and $P_{GC}^{(2)} = 0.189$. The approach to the asymptotic values of G, A, C, and T of 0.27, 0.28, 0.17, and 0.28 is thus not all that rapid, and the latter cannot replace the exact calculations. Further, assuming the idealized case in Example 1 where any base is equally likely to mutate to and be fixed as any of the other three, the expected base composition is for G, A, T, and C, respectively 0.333, 0.222, 0.222, and 0.222 (Holmquist, 1972). The relative error introduced by the idealization of Example 1, averages about 27%, ranging from 12% for G to 55% for T and is clearly not negligible.

Example 3. As an extension of Example 2, consider the probability that after 10 fixed nucleotide point mutations the triplet (3') GAC (5') will end up TTC. Let the 10 mutations be distributed over the first, second, and third nucleotide positions of the triplet, respectively, 3,5, and 2, and assume the nucleotide transition probabilities are given at each of the three nucleotide positions by the matrix of Example 2. This process represents one pathway by which leucine (codon CUG) could be

converted to lysine (codon AAG) in a protein. Compare the results with that obtained unter the idealized assumptions of Example 1.

We calculate the idealized case first. From Table 1 in Holmquist (1972) the probability of GAC → TTC by the described pathway is:

$$\{[1/3][1 - (2/9)]\}\{[1/3][1 - 20/81]\}\{1/3\} = 0.0217.$$

From example 2,

$$P_{GT}^{(3)} = 0.1667 - \frac{0.1245(-1)^3}{(6.010)^3} + \frac{0.5827}{(2.395)^3}\cos[4.640+3.080(3)] = 0.1781.$$

The calculations for $P_{AT}^{(X)}$ and $P_{CC}^{(X)}$ are analogous to those already given in Example 2 with the results:

$$P_{AT}^{(5)} = 0.1667 - \frac{0.002379}{(6.010)^5}(-1)^5 + \frac{0.2703}{(2.395)^5}\cos[2.224+3.080(5)] = 0.1678.$$

$$P_{CC}^{(2)} = 0.2783 + \frac{0.2024}{(6.010)^2}(-1)^2 + \frac{1.209}{(2.395)^2}\cos[1.127+3.080(2)] = 0.3973.$$

The correct probability of GAC → TTC by the described pathway is thus:

$$P_{GAC\to TTC}^{(3,5,2)} = P_{GT}^{(3)} P_{AT}^{(5)} P_{CC}^{(2)} = (0.1781)(0.1678)(0.3973) = 0.0119.$$

The relative error committed by using the approximations of Example 1 is thus quite large, 82%, i.e. $[100(0.0217 - 0.0119)/0.0119]$.

Example 4. Estimates of the twelve transition probabilities can be obtained by examining adjacent ancestral nodal sequences in phylogenetic trees which have been constructed by the method of parsimony and tabulating for all such adjacent nodes the number of times each base in the most ancestral of the two adjacent nodes is replaced by the other bases in the more recent of the two nodes. Goodman and his colleagues (1974) have made a partial tabulation of this sort for the first two codon positions along six lineages of metazoan globin chains. A total of 867 base fixations were tabulated. These data have been analyzed (Holmquist, 1976) with the following results, averaged for the six lineages and two codon positions:

	T	G	C	A
T	0	0.44	0.40	0.16
G	0.42	0	0.37	0.21
C	0.49	0.32	0	0.19
A	0.34	0.37	0.39	0

The magnitudes of the deviations of these matrix elements from the idealized values of 1/3 are generally similar to those of the hypothetical matrix in Example 2, but values for the magnitudes themselves differ from those in the example. For comparative purposes it is instructive to compute the same quantities as for Examples 2 and 3. The calculations proceed analogously to those already given, but more simply, as in the present case D(s) has no complex zeros. The results are:

$$P_{GG}^{(k)} = 0.2756 + \left(\frac{0.0014}{5.656^k} + \frac{0.0606}{2.230^k} + \frac{0.6623}{2.668^k}\right)(-1)^k$$

$$P_{GA}^{(k)} = 0.1567 - \left(\frac{0.0320}{5.656^k} + \frac{0.0117}{2.230^k} + \frac{0.1131}{2.668^k}\right)(-1)^k$$

$$P_{GT}^{(k)} = 0.3001 + \left(\frac{0.0150}{5.656^k} - \frac{0.0601}{2.230^k} - \frac{0.2550}{2.668^k}\right)(-1)^k$$

$$P_{GC}^{(k)} = 0.2675 + \left(\frac{0.0156}{5.656^k} + \frac{0.0111}{2.230^k} - \frac{0.2941}{2.668^k}\right)(-1)^k .$$

Again taking k = 2, the expected base composition after two base replacements at the site initially occupied by G is, in the order G, A, T, and C, 0.381, 0.137, 0.253, and 0.229. The asymptotic values cannot replace the more exact calculations. In this (k = 2) example the idealization used in Example 1 would introduce an average relative error in base composition of 22% ranging from 3% for C to 62% for A.

Finally

$$P_{GAC\rightarrow TTC}^{(3,5,2)} = P_{GT}^{(3)} P_{AT}^{(5)} P_{CC}^{(2)} = (0.3189)(0.3003)(0.3695) = 0.0354.$$

The relative error committed by using the approximations of Example 1 is 39% and not negligible.

For longer gene sequences representative of the structural genes of proteins (as contrasted with the simple nucleotide triplet considered here) and for other protein families the magnitude of the deviations from the idealized values (as in Example 1) of the probability with which one DNA sequence goes to another may be greater or less than in this example. For more accurate calculations, one should use a separate transition matrix for each codon position. The true average deviation of the matrix elements from the equiprobable values of 1/3, on a per codon basis, must be less than that calculated here because the third nucleotide position of each codon is known from the experiments of Salser and his co-workers (1976) to be under less selective constraint than the first two positions.

CONCLUSION

The purpose of this paper has been to derive usable analytical expressions which can accurately describe the dynamics of the evolutionary divergence of two genes when the distribution of fixed point mutations along the gene and the twelve transition probabilities of the four bases to each other are known. It has been of equal importance to express these equations directly in terms of biologically familiar interpretable parameters - the twelve transition probabilities of one nucleotide to another - and to provide simple numerical illustrative examples so that calculations of this type are accessible to practicing molecular evolutionists and not only to mathematicians. The assumption that each base is equally likely to mutate to and be fixed as any one of the other three is shown to lead to appreciable error in realistic biological applications. These equations can also help in reducing the necessity for expensive Monte Carlo type computer simulations.

When applying these methods to genes coding for protein structure, some of the mutational pathways involved may include one or more nucleotide triplets which code for one of the three chain-terminating codons UUA, UAG, or UGA in the messenger RNA. Each investigator must allow for this particularity in a manner consistent with the application in mind.

The full potential of this work can be realized only after calculations for long gene sequences. It is now clear, however, that such calculations are both theoretically tractable and practically implementable without undue demands of time. Such calculations are a necessary prerequisite for an understanding of the evolution of macromolecules.

Acknowledgement. This work was supported by grant NGR 05-003-460 from the National Aeronautics and Space Administration. I thank Professor Charles Antoniak in the Department of Statistics at the University of California at Berkeley for critically reading this paper.

REFERENCES

D'Azzo, J.J., Houpis, C.H. (1966). In: Feedback control system analysis and synthesis, 2nd Ed., Chapter 4: Laplace transforms, Section 7, Heaviside partial-fraction expansion theorems, pp. 94-115. New York: McGraw-Hill

Feller, W. (1968). Integral-valued variables. Generating functions. In: An introduction to probability theory and its applications, Volume I, 2nd Ed., Chapter XI, pp. 264-276. New York: Wiley

Goodman, M., Moore, G., Barnabas, J., Matsuda, G. (1974). J.Mol.Evol. 3, 1

Holmquist, R. (1972). J.Mol.Evol. 1, 115 (see particularly Table 1, on page 122)

Holmquist, R. (1976). Random and non-random processes in the molecular evolution of higher organisms. In: Molecular anthropology, M. Goodman, R.E. Tashien, eds., Table IV. New York: Plenum Press

Holmquist, R., Jukes, T.H., Moise, H., Goodman, M., Moore, G. (1976). J.Mol.Biol. 105, 39

Jukes, T.H., Holmquist, R. (1972). J.Mol.Biol. 64, 163

Keller, E.G., Doherty, R.E. (1961). Applications of determinants. In: Mathematics of modern engineering, Volume I, Chapter II, 63. New York: Dover

Moore, G., Goodman, M., Callahan, C., Holmquist, R., Moise, H. (1976). J.Mol.Biol. 105, 15

Salser, W., Bowen, S., Browne, D., El Adli, F., Fedoroff, N., Fry, K., Heindell, H., Paddock, G., Poon, R., Wallace, B., Whitcome, P. (1976). Fed.Proc.Fed.Am.Soc.Exp.Bul. 35, 23