J Mol Evol (1985) 22:69–81

Journal of
Molecular Evolution
© Springer-Verlag 1985

A Method of Estimating from Two Aligned Present-Day DNA Sequences Their Ancestral Composition and Subsequent Rates of Substitution, Possibly Different in the Two Lineages, Corrected for Multiple and Parallel Substitutions at the Same Site

B. Edwin Blaisdell

Linus Pauling Institute of Science and Medicine, Palo Alto, California 94305, USA

Summary. The course of evolutionary change in DNA sequences has been modeled as a Markov process. The Markov process was represented by discrete time matrix methods. The parameters of the Markov transition matrices were estimated by least-squares direct-search optimization of the fit of the calculated divergence matrix to that observed for two aligned sequences. The Markov process corrected for multiple and parallel substitutions of bases at the same site. The method avoided the incorrect assumption of all previously described methods that the divergence between two present-day sequences is twice the divergence of either from the common and unknown ancestral sequence. The three previous methods were shown to be equivalent. The present method also avoided the undesirable assumptions that sequence composition has not changed with time and that the substitution rates in the two descendant lineages were the same. It permitted simultaneous estimation of ancestral sequence composition and, if applicable, of different substitution rates for the two descendant lineages, provided the total number of estimated parameters was less than 16. Properties of the Markov chain were discussed. It was proved for symmetric substitution matrices that all elements of the equilibrium divergence matrix equal $1/16$, and that the total difference in the divergence matrix at epoch k equals the total change in the common substitution matrix at epoch 2k for all values of k. It was shown how to resolve an ambiguity in the assignment of two different substitution rates to the two descendant lineages when four or more similar sequences are available. The method was applied to the divergence matrix for codon site 3 for the mouse and rabbit beta-globins. This observed divergence matrix was significantly asymmetric and required at least two different substitution rates. This result could be achieved only by using different asymmetric substitution matrices for the two lineages.

Key words: Evolutionary change — Markov process — Discrete matrix model — Direct-search optimization — Substitution matrices — Equilibrium divergence matrix symmetry — Mouse beta-globin — Rabbit beta-globin

Introduction

Zuckerkandl and Pauling (1962) first suggested that the evolutionary distance between two protein (or DNA) sequences X and Y can be inferred from the observed divergence matrix of counts of the occurrences of amino acids (bases) i_x and j_y, $1 \le i_x, j_y \le 20$ $(1 \le i_x, j_y \le 4)$, in corresponding sites in the sequences. They used the raw differences, but pointed out that these should be corrected for multiple substitutions at the same site. Later (Zuckerkandl and Pauling 1965), they showed that in long sequences the fraction of unsubstituted elements declines exponentially with time. They also discussed at length the evidence that the rate of substitution of one amino acid by another depends on such properties of the two amino acids involved as bulk, charge, polarity, and ability to interact with other amino acids in the sequence to produce a functioning three-

Offprint requests to: B.E. Blaisdell

Table 1. Substitution-matrix models

Model[a]		T	C	A	G	Sum	%
I	T		a	a	a		
	C	a		a	a		
	A	a	a		a		
	G	a	a	a			
II	T		a	b	b		
	C	a		b	b		
	A	b	b		a		
	G	b	b	a			
III	T		a	b	c		
	C	a		c	b		
	A	b	c		a		
	G	c	b	a			
IV	T		a	b	da		
	C	c		dc	b		
	A	b	da		a		
	G	dc	b	c			
V	T		a	b	a		
	C	c		c	d		
	A	e	a		a		
	G	c	f	c			
VI	T		a	b	c		
	C	a		d	e		
	A	b	d		f		
	G	c	e	f			
VII	T		a	b	b		
	C	c		b	b		
	A	d	d		a		
	G	d	d	c			
VIII	T			a	b		
	C			c	d		
	A	e	f				
	G	g	h				
RM	T	25	11	1	1	38	27
	C	5	33	1	0	39	28
	A	3	1	2	4	10	7
	G	5	5	4	40	54	38
	Sum	38	50	8	45	141	
	%	27	35	6	32		
SRM	T	25	6	2	2	35.25	25
	C	6	33	2	2	43.25	31
	A	2	2	2	6	12.25	9
	G	2	2	6	40	50.25	36

[a] I, Jukes and Cantor (1969). II, Kimura (1980). III, Kimura (1981). IV, Takahata and Kimura (1981). V, Gojobori et al. (1982). VI, Lanave et al. (1984). VII, Blaisdell (present paper). VIII, Holmquist (1976). RM, observed rabbit–mouse beta-globin divergence matrix. SRM, symmetrized two-parameter rabbit–mouse divergence matrix; correct transversion values are 2.125. For matrices I–VII each element on the main diagonal is 1 minus the sum of the other elements in its row. In matrix VIII, each element on the main diagonal is zero and the other four unprinted elements are 1 minus the sum of the other elements in their respective rows

dimensional protein structure. Presumably, a similar variability of rate of substitution would be found in the base sequences that generate the protein sequences.

Jukes and Cantor (1969) provided a formula for correcting the total number of substitutions in a divergence matrix for multiple substitutions at the same site. They assumed that all kinds of base substitutions were equally likely (Table 1, substitution matrix I). They did not disclose their method of arriving at their formula. Kimura (1980) derived their formula by using systems of linear differential equations with constant coefficients as an expression of the Markov process (Feller 1968, p. 444). His derivation showed that their formula depends on assuming that the rates of substitution in the two sequences are equal, that the difference between the two strands is twice the difference of each from the common ancestor, and that base composition does not change with time. He obtained similar results to theirs for a rate matrix (Table 1, matrix II) having one rate for transitions (purine–purine or pyrimidine–pyrimidine substitutions) and a second for transversions (purine–pyrimidine substitutions), a two-parameter model. Note that his rate matrix is symmetric; that is, the rate of substitution of base i by base j equals the rate of substitution of base j by base i. This treatment of the divergence matrix was probably motivated by the observation of Fitch (1980) that in three beta-globins, among fourfold degenerate codons, even third-position transitions outnumber transversions by 2 to 1.

Similar differential equation solutions have been given for divergence matrices with three (Kimura 1981), four (Takahata and Kimura 1981), and six parameters (Gojobori et al. 1982) (Table 1, matrices III, IV, and V, respectively). Note that matrices IV and V are not symmetrical. These solutions are closed algebraic expressions the values of whose parameters are determined from imposition on observed divergence matrices of the structures of the various models given in Table 1. Tajima and Nei (1984) recognized that actual divergence matrices may be unlike any of the restricted structures of Table 1, matrices I–V, and suggested an ad hoc approximation that gave fairly good (error less than 10%) estimates of the number of substitutions if this number was small (less than one per base on the average).

Lanave et al. (1984) used the solutions of the differential equations corresponding to the six-parameter symmetrized observed divergence matrix (Table 1, matrix VI) and determined their dispersion by Monte Carlo computer simulation. This method is the most flexible of the differential-equation methods and makes fewer artificial assumptions. (In Table 1, matrices I–VII, each element on the main diagonal is 1 minus the sum of the other elements in its row.)

Holmquist (1976) gave numerical procedures for solving by generating functions the Markov process defined by relative substitution rates (eight parameters) at a single base position (Feller 1968, p. 264). In his matrix (Table 1, matrix VIII), the elements on the main diagonal are zero and the sum of the three rates in each row is 1; that is, the elements in matrix VIII are the probabilities of a mutation conditioned on there being a mutation for a given base. These solutions give the exact probabilities for each base substitution for an arbitrary number of mutations. When supplemented with the assumption of a Poisson distribution for the number of mutations at a given site, these procedures give the same results as those of Lanave et al. and those of Kimura et al. when applied to the same model (Table 2).

I have approximated the Markov process by discrete time matrix methods that permit the estimation of substitution rates that may be different for the two descendant lineages and the simultaneous estimation of the ancestral base composition. This is done by minimizing the sum of squares of the differences between the predicted and observed divergence matrices, as in Table 1, matrix RM. Matrix RM is data for the mouse and rabbit beta-globin sequences: The first row gives the counts of occurrences of T, C, A, and G in the mouse at sites corresponding to the occurrence of T in the rabbit; the first column gives the counts of T, C, A, and G in the rabbit at sites corresponding to the occurrence of T in the mouse; and similarly for the other rows and columns. This method makes unnecessary the assumptions common to all the other methods, namely that composition does not change with time, that the rates of substitution for the two lineages are equal, and that the present difference between the two sequences is twice the difference between each and the common ancestor. In common with the other methods, it does assume that the process of evolutionary base substitution may be modeled by a Markov process, that the rates of substitution are the same at all sites in the sequence, and that the rates of substitution are constant over time (a stationary Markov process). More precisely, since I treat the Markov process as a Markov chain with a finite number of states (four for DNA) and its development through successive epochs by matrix multiplication, the assumption of constant probability of substitution per epoch does not entail that epochs correspond to constant intervals of time. The unreality of the assumption that the rates of substitution are the same at all sites can be ameliorated, as in the example below, by considering only the "silent" codon sites 3, and especially so if only those in which all four bases in site 3 code for the same amino acid are considered. In particular, I have considered in this paper Table 1, matrix VII, which maintains the distinction between rates of transition and transversion, and permits them to be different in the forward and backward directions. From this it follows that the equilibrium counts of pyrimidines (or purines) need not be the same and that the counts of base i occurring in gene X at sites where base j occurs in gene Y need not be the same as the counts of base j occurring in gene X where base i occurs in gene Y.

Table 2 Comparison of Holmquist (1976), Kimura (1980), and Lanave et al. (1984) two-parameter solutions, *P* (transition) = *P* (transversion)

A	0	1	0	0	606,531
	1	0	$\frac{1}{2}$	$\frac{1}{4}$	303,265
	2	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	75,816
	3	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{1}{4}$	12,636
	4	$\frac{9}{32}$	$\frac{7}{32}$	$\frac{1}{4}$	1,580
	5	$\frac{15}{64}$	$\frac{17}{64}$	$\frac{1}{4}$	158
	6	$\frac{33}{128}$	$\frac{31}{128}$	$\frac{1}{4}$	13
	7	$\frac{63}{256}$	$\frac{65}{256}$	$\frac{1}{4}$	1
	8	$\frac{129}{512}$	$\frac{127}{512}$	$\frac{1}{4}$	0
B	O = 637,816 P = 165,499 Q = 98,367				
C	$-(\frac{1}{2})\ln[(1 - 2P - 2Q)\sqrt{1 - 4Q} = \frac{1}{2}$				
D	$-(\frac{3}{4})\ln[1 - 4(P + 2Q)/3] = 494{,}657$				
E		637,816	165,449	98,367	98,367
		165,449	637,816	98,367	98,367
		98,367	98,367	637,816	165,449
		98,367	98,367	165,449	637,816
F		472,367	472,367	606,531	1,000,000
G		$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{2}$	0
H		0	$-1/\sqrt{2}$	$\frac{1}{2}$	$\frac{1}{2}$
		0	$1\sqrt{2}$	$\frac{1}{2}$	$\frac{1}{2}$
		$-1/\sqrt{2}$	0	$-\frac{1}{2}$	$\frac{1}{2}$
		$1/\sqrt{2}$	0	$-\frac{1}{2}$	$\frac{1}{2}$
I	$(\frac{1}{4})[$	0	+ 0	$+(\frac{1}{2})(\frac{3}{4})$	$+(\frac{1}{2})(\frac{3}{4})$
		$+(\frac{1}{2})(\frac{3}{4})$	$+(\frac{1}{2})(\frac{3}{4})$	+ 0	+ 0
		$+(\frac{1}{4})(\frac{1}{2})$	$+(\frac{1}{4})(\frac{1}{2})$	$+(\frac{1}{4})(\frac{1}{2})$	$+(\frac{1}{4})(\frac{1}{2})]$
	$= \frac{1}{2}$				

All decimal values have been multiplied by 1,000,000. A, Holmquist solution: col. 1, numbers of substitutions per site; col. 2, P (unchanged); col. 3, P (transition); col. 4, ½ P (transversion); col. 5, Poisson probability density for mean P = ½. B, Poisson averaged values: O, P (unchanged); P, P (transition); Q, ½P (transversion). C, Kimura two-parameter solution. D, Jukes and Cantor (1969) one-parameter solution. E, Substitution matrix (T, C, A, G arranged as in Table 1, matrix II). F, Eigenvalues of row E values. G, Negative of natural logarithms of row F values. H, Eigenvectors of row E values. I, Lanave et al. solution (assuming fraction of each base = ¼): sum[ln(eigenvalue) × square (eigenvector values)] = ½

Description of the Discrete Matrix Optimization Method

Assume the substitutions in the descent of one lineage from the common ancestor are modeled by the first-order Markov transition matrix S = s(i, j), i, j = 1, 2, 3, 4, where s(i, j) is the probability that the base at a given site is j at epoch k conditioned on its being i at epoch k − 1 and independent of its values at epochs k = 0, 1, 2, . . . , k − 2. Then S(k) = s(i, j, k) = S(1)^k, where s(i, j, k) is the probability that the base is j at epoch k conditioned on its being i at epoch 1. Similarly, let T(k) be the transition matrix for the second lineage. Then the divergence probability matrix for the divergence between the two lineages at epoch k is

$$D(k) = d(i, j, k) = \sum c(l, 0)s(l, i, k)t(l, j, k) \quad (1)$$

where i, j, l = 1, 2, 3, 4; k = 1, 2, 3, . . . ; c(l, 0) is the fraction of base l in the ancestral sequence at epoch 0; and d(i, j, k) is the fraction at epoch k of the number of occurrences at the same site in the two descendant sequences of j in the second sequence and i in the first sequence. Note that the two sequences are distinguished so that d(i, j, k) may not equal d$^{i,j}_{j}$(j, i, k). Note also that $\sum_{i,j}$ d(i, j, k) = 1, so there are 15 independent values in D(k). Each of the rows of S and T sums to 1, so there are 12 independent values in each, and there are three independent values in ancestral composition c, making a total of 27 parameters determining the course of the evolutionary substitutions. Obviously, 27 parameters cannot be estimated unambiguously from 15 independent observations. The number of estimable parameters is reduced to 15 or fewer by imposing a reasonable structure on the matrices S and T; for example, if it is assumed in matrix II of Table 1 that the transition rate equals 4a and the transversion rate equals 8b, there are five estimable parameters: a, b, and three ancestral compositions, c(1, 0), c(2, 0), and c(3, 0). The parameters are estimated by nonlinear least-squares minimizing of the squares of the differences between the calculated divergence matrix (Eq. 1) and the observed one (e.g., Table 1, matrix RM) using the iterative "complex" optimization of Box (1965) to determine a, b, c(1, 0), c(2, 0), and c(3, 0). In the use of Eq. (1), one must choose a reasonable number of epochs. From Table 6 it appears that 16 is large enough to yield agreement within 1% with a very large number. In the application in Table 9, I used four epochs to speed calculation.

The optimizer can be given initial values derived from the observed divergence matrix as follows: For initial composition, use the average values for the two genes in the observed divergence matrix. For example, for Table 1, matrix RM, I used (0.28 + 0.35)/2 = 0.315 for C and similarly for T and A. Then G = 1 − (T + C + A). (Here each letter denotes the composition fraction of the corresponding base.) For initial rates, use average values from the observed divergence matrix divided by the chosen number of epochs. For example, in Table 1, matrices II and RM, for the transition rate b I used (1 + 1 + 1 + 0 + 3 + 1 + 5 + 5)/(141 × 4 × 2) = 0.015071. These initial values will be too large, because, for example, for transitions the calculation of the rate of simultaneous occurrence of T and C ignores the possibility that in addition to one arising from the other, both may have arisen from a purine (A or G). I have also used the simplex method of Spendley et al. (1962) and the quadratic method of Powell (1964), but have found the complex method generally to be faster and less prone to failure. These are all direct search methods, because closed expressions for the elements of the divergence matrix for large epochs k are practically unattainable. The minimization is nonlinear because even the elements of the square of a matrix are second degree in the elements of the original matrix. The speed and reliability of the optimizations decline rapidly with an increase in the number of estimated parameters, so it is desirable to keep this number small. Guidance in the selection of a model is discussed below.

Comparison of the Three Earlier Methods

The three earlier solutions, (1) differential equation solution in closed form for a specific substitution matrix (Kimura 1980), (2) differential equation solution of the symmetrized observed divergence matrix (Lanave et al. 1984), and (3) the generating function solution (Holmquist 1976), supplemented with the assumption that the number of substitutions at a given site is Poisson distributed, all give the same result for the same model. This is shown in Table 2 for the two-parameter model (Table 1, matrix II). Rows A display the generating function solution (Holmquist 1976). The first column gives the numbers of mutations per site; the second, the probability that the occupant of a site is unchanged; the third, the probability that a site shows a transition substitution; the fourth, half the probability that a site shows a transversion substitution; and the fifth, the probability mass values for a Poisson distribution of mean value ½. In row B, O is the expectation that a site shows no change, and is the inner product of the second and fifth columns in rows A. Similarly, P is the expectation that a site shows a transition and Q is half the expectation that the site shows a transversion. Row C shows that application of the closed-form differential equation solution of Kimura (1980) for the two-parameter model (Table 1, matrix II) recovers the correct rate of substitution of ½. Row D shows that the misapplication of the closed form for the one-parameter model (Table 1, matrix I) leads to an underestimation by 1%. The calculation of the correct value in row C shows that the closed-form differential equation solution and the generating function solution supplemented with

Table 3. Comparison of fraction changed and fraction diverged difference for one-parameter solutions

A	B	C	D	E	F	G	H	I
1/256	3,906	3,896	3,906	7,772	7,771		7,791	
1/128	7,812	7,772	7,812	15,463	15,461		15,501	
1/64	15,625	15,463	15,625	30,608	30,604	30,607	30,682	30,925
1/32	31,250	30,608	31,250	59,967	59,959	59,965	60,108	60,575
1/16	62,500	59,967	62,500	115,139	115,122	115,135	115,398	116,258
1/8	125,000	115,139	125,000	212,602	212,568	212,593	213,036	214,494
1/4	250,000	212,602	250,000	364,937	364,866	364,920	365,546	367,646
1/2	500,000	364,937	500,000	552,302	552,151	552,265	552,880	555,078
1/1	1,000,000	552,303	1,000,000	697,887	697,571	697,808	698,050	699,352

All decimal values have been multiplied by 1,000,000. A, Rate of substitution; B, decimal equivalents of values in A; C, fraction of bases changed; D, Jukes–Cantor one-parameter calculation from column C of fraction changed; E, fraction of two diverged sequences that differs, calculated from column C using equation of Holmquist (1972) [= 2C − (4/3)C^2]; F, discrete time matrix calculation of fraction differing using fraction changed at substitution rate = 1/256; G, same as F, but substitution rate = 1/64; H, discrete time matrix calculation of fraction differing using substitution rate = 1/256; I, same as H, but substitution rate = 1/64

the Poisson distribution for the number of mutations give the same result. Rows E through I outline the steps of the numerical differential equation solution for the symmetrized substitution matrix (rows E), which is assumed to be the same as the observable divergence matrix (Lanave et al. 1984). Row F gives the eigenvalues of the matrix in rows E, and row G shows the negative natural logarithms of the row F values. The columns of rows H are the eigenvectors corresponding to the eigenvalues. Rows I illustrate the calculation of the average substitution rate [Eq. (20) in Lanave et al. (1984)],

$$\text{average rate} = \sum_{i,j} q(i)[H(i, j)]^2 G(j)$$

which also recovers the correct value of ½. (Here G and H denote values in the corresponding parts of Table 2.) In rows I the columns correspond to base indices i = 1, 2, 3, 4, and the rows to eigenvalues (and eigenvectors) j = 1, 2, 3. There is no row for j = 4, since G(4) = 0. The calculation is shown for a value of q(i), the fraction of base i in the ancestral sequence, equal to ¼ for all i. However, the same value would be obtained for an arbitrary set of q(i), since these sum to 1 by definition and the sum of each column in I is the same. This completes, for the example of Table 1, matrix II, the demonstration that the three earlier solutions give the same results for the same model.

Demonstration That the Three Earlier Methods Do Not Give the Correct Values for the Divergence Matrix, Whereas the Discrete Time Method Does

The three earlier methods discussed in the preceding section give the correct result only for the evolutionary distance between the observed present-day sequence and its inaccessible ancestral sequence. When using the present-day divergence between two sequences to infer the evolutionary distance between them, all these methods make the obviously incorrect assumption that the divergence between two sequences is twice the divergence of each from their common ancestral sequence. It is clear that the divergence between two sequences will on the average be less than twice the divergence of each from the common ancestral sequence, because some of the net substitutions in the two lineages may be the same. The consequences of this incorrect assumption are shown in Tables 3 and 4.

For ease of reference, I shall from now on call the matrix of differences between a present-day sequence and its ancestral sequence a substitution matrix and the matrix of differences between two present-day sequences a divergence matrix. Holmquist (1972) has given for the one-parameter model (Table 1, matrix I) a closed expression for calculating the fraction in the divergence matrix different from the net fraction changed in the substitution matrix. The effect of this correction is shown in Table 3. Columns A and B show the mean values for the number of substitutions in an assumed Poisson distribution. Column C shows the average net number of substitutions calculated for the one-parameter model as Table 2, row B, was calculated for the two-parameter model (see description above). Column C shows very clearly the effect of multiple mutations. When the number of mutations per site is 1.00, only 0.55 substitutions will be observed. Column D is calculated from the observable column C using the closed-form differential equation solution (Jukes and Cantor 1969) and the correct numbers of substitutions are recovered; i.e., column D is the same as column B. Column E shows the divergence values calculated from the substitution values in column C using Holmquist's (1972) correction expression. The observed divergence values are much less than twice the substitution values at the higher numbers of substitutions. When the number of substitutions per site is 0.500, the observed number of substitutions is 0.365. Twice this value is 0.710, but the observed divergence is only 0.552. In fact, when the number of substitutions per site is 1.000, the observed substitution number is 0.552, and twice this value is 1.104, which is impossible, because the observed fraction of substitutions cannot be greater than 1. In fact, as the average number of mutations becomes very large, both the observed number of substitutions (column C) and the observed divergence (column E) approach ¾.

Columns F and G compare the discrete time matrix calculation of the observed divergence values with the generating function values (column E), the latter of which are equal to the closed-form differential equation (infinitesimal) values, as shown above. Columns F and G are started in the first row with the observable average fractions changed taken from column C at substitution rates of 1/256 and 1/64, respectively, and values in lower rows are obtained by matrix multiplication. The small discrepancies between the values in F and G and the more accurate values in column E are undoubtedly attributable to accumulation of round-off error during the multiple single-precision matrix multiplications. The difference between the epoch sizes of columns F and G does not seem to be important (see also Tables 4 and 5, below). In practice it is desired to estimate the true substitution rate (column A) from the observed divergence. Columns H and I are started with the true rates 1/256 and 1/64. The values in columns H and I differ from the more accurate values in column E by larger systematic amounts, but they extrapolate easily at rate zero to the values in column E (data not shown).

Note that in Table 3, column E, row i, is the same as column C, row i + 1, for all i; that is, in this one-parameter example, divergence at k equals substitution at 2k, where k is the epoch, and is not equal to twice substitution at k, as is assumed incorrectly in the differential equation solutions. This means that application of the Jukes–Cantor one-parameter formula to the observed divergence matrix gives twice the rate of substitution in each lineage. It may be argued that in the inference of an evolutionary tree from an observed divergence matrix a factor of 2 in all evolutionary distances would not alter the tree. This is correct for the one-parameter model.

However, although it is proved below that total fraction difference in divergence at k is the same as total fraction changed in substitution at 2k for all symmetric substitution matrices, in models depending on more than one parameter, no such simple relation holds for the various classes of base change. This is shown for the two-parameter transition–transversion model in Table 4, columns C and D. In Table 4 column A shows the epoch, k. Column B shows the substitution probability at epoch 2k, S(2k), calculated by matrix multiplication of the two-parameter model (Table 1, matrix II) with the transition probability a = 2/1024 and half the transversion probability b = 1/1024, or a total probability of change of 1/256. The last row of Table 4 shows one result for a total probability of change of 1/64. Columns C and D show, respectively, the probabilities at epoch k of a transition difference and of a transversion difference as found in D(k), which is calculated from S(k) as described above. Column E shows the sum from columns C and D of the probabilities of a divergence difference at epoch k, which in this symmetric two-parameter model is equal also to the probability of substitution at epoch 2k. The small discrepancy is undoubtedly attributable to accumulation of round-off error during the multiple single-precision matrix multiplications. In fact, if the two values are derived from the much larger main diagonal elements of S(2k) and D(k), the discrepancy is less than 1/10,000 in the worst case, epoch 256 (data now shown). Column F is the probability of substitution calculated from columns C and D by means of the closed-form differential equation solution (Kimura 1980):

$$-\tfrac{1}{2}\ln[(1 - 2C - D)\sqrt{1 - 2D}]$$

(Here the letters denote the values in the corresponding columns.) For low numbers of epochs, the values in F are nearly equal to those in B (see their ratio in column G), but the error increases to nearly three-fold at epoch 256. Also recall that the values calculated from the divergence matrix at epoch k, which approximate at low k the values in the substitution matrix at epoch 2k, are approximately equal to twice the values in the substitution matrix at epoch k, as found exactly for the one-parameter model in Table 3. The values in the last row for a substitution probability of 1/64 at epoch 4 are approximately equal to the values for a substitution probability of 1/256 at epoch 16, which shows that the graininess of the discrete time matrix method is of little moment.

Properties of the Discrete Time Matrix Solution

Properties of the discrete time matrix solution are summarized in Table 5. Properties are listed for four classes of solutions according to whether the substitution matrices for the two descendant lineages are the same or different and whether for each of these the substitution matrix is symmetric or asymmetric. I call attention to the following features. The substitution matrices show that their Markov chains are regular, that is, that for large numbers of epochs

Table 4. Comparison of the discrete time matrix solution with the closed-form differential equation solution for the two-parameter model

A	B	C	D	E	F	G
			Substitution probability = $\frac{1}{256}$ = 0.003906			
1	7,791	3,892	3,900	7,792	7,834	1.0055
2	15,499	7,731	7,770	15,501	15,668	1.0109
4	30,667	15,252	15,418	30,671	31,336	1.0218
8	60,042	29,688	30,362	60,050	62,673	1.0438
16	115,136	56,272	58,880	115,152	125,347	1.0887
32	212,115	101,320	110,828	212,148	250,698	1.1819
64	362,825	165,798	197,098	362,896	501,424	1.3820
128	547,080	230,706	316,526	547,232	1,003,033	1.8334
256	691,506	259,066	432,754	691,820	2,008,373	2.9043
			Substitution probability = $\frac{1}{64}$ = 0.015625			
4	116,003	56,811	59,190	116,001	126,360	1.0893

Columns A–F have been multiplied by 1,000,000. A, Epoch (k); B, probability of substitution at epoch 2k; C, P (= probability of transition divergence difference at epoch k); D, Q (= probability of transversion divergence difference at epoch k); E, (= probability of divergence difference at epoch k); F, probability of substitution calculated by Kimura (1980) two-parameter model (= $-\frac{1}{2}\ln[1 - 2P - Q)\sqrt{1 - 2Q}]$); G, (value in F)/(value in B)

Table 5. Properties of the discrete matrix solutions

		Substitution matrices for the two lineages			
		Equal		Unequal	
		Symmetric	Asymmetric	Symmetric	Asymmetric
A	Equilibrium substitution matrices, i, j = 1, 2, 3, 4, c(j, e) = equilibrium composition				
	s(i, j)	¼	c(j, e)	¼	c(j, e)
	t(i, j)			¼	c(j, e)
B	Equilibrium divergence matrix d(i, j), row sums = 1, x(j, e) = equilibrium differences	¼	x(j, e)	¼	x(j, e)
C	Equilibrium divergence matrix e(i, j), sum of all elements = 1	$\frac{1}{16}$	Sym.	$\frac{1}{16}$	Asym.
D	Equilibrium substitution composition preserved at all epochs k?				
	s(i, j)	Yes	Yes	Yes	Yes
	t(i, j)			Yes	Yes
E	Ancestral substitution composition approaches equilibrium composition?				
	s(i, j)	Yes	Yes	Yes	Yes
	t(i, j)			Yes	Yes
F	Equilibrium divergence differences preserved for all epochs k?	Yes	No	Yes	No
G	Early divergence differences approach equilibrium differences?	Yes	Yes	Yes	Yes
H	Divergence difference at k = substitution change at 2k?	Yes	No	Yes	No

k each base can be substituted by every base (i.e., there are no zeros in the substitution matrix for large numbers of epochs k). Markov chain theory proves that for large k values the substitution matrix approaches an equilibrium matrix in which all rows are the same and are equal to the equilibrium composition (rows A) and that if the chain process is started at the equilibrium composition it will stay at it (row D), but that if it is started at any other composition it will of course approach the equilibrium composition (row E).

Kimura (1980) noted that for the two-parameter solution, at equilibrium P = Q/2 = ¼ even when the transition probability is not equal to half the trans-

Table 6. Determination of substitution rate–epoch pairs for the two-parameter model

A	B	C	D	E	F	G
8	11,559	5,814	3,853	1,938	92,472	35
12	7,759	3,891	3,880	1,946	93,108	19
16	5,839	2,924	3,893	1,949	93,424	50
24	3,906	1,953	3,906	1,953	93,744	28
32	2,935	1,466	3,913	1,955	93,920	64
48	1,960	978	3,920	1,956	94,080	17
64	1,471	734	3,923	1,957	94,144	39

Columns B–F have been multiplied by 1,000,000; column G, by 1,000,000,000. A, Epoch. B, Transition rate determined by best fit to observed divergence matrix at epoch 24; values for epoch 24 are correct. C, Same as B for ½× transversion rate. D, Transition rate calculated for epoch 24 by ratio of epochs. E, Same as D for ½× transversion rate. F, A(B + 2C)/2 (letters denote values in corresponding columns). G, Sum of squares of residuals of fit of calculated best-fit divergence matrix to observed divergence matrix

version probability. This can be proved in a more general way for any symmetric matrix (rows A). Let S(1) = s(i, j, l) at epoch 1, where s(i, j, l) = s(j, i, l) by symmetry. Then at epoch 2, $S(2) = S^2(1) =$ s(i, j, 2), where s(i, j, 2) = $\sum_k$ s(i, k, l)s(k, j, l) = $\sum_k$ s(k, i, l)s(j, k, l) = s(j, i, 2), or S(2) is symmetric, since s(i, k, l) = s(k, i, l) and s(k, j, l) = s(j, k, l). Similarly, S(k) is symmetric for all k. Therefore, at equilibrium s(1, 2) = s(2, 1), s(1, 3) = s(3, 1), and s(1, 4) = s(4, 1), and therefore s(1, 1) = s(1, 2) = s(1, 3) = s(1, 4), since s(1, 1) = s(2, 1) = s(3, 1) = s(4, 1) at equilibrium. Since S is a probability matrix, each row sums to 1. Thus s(1, 1) = s(1, 2) = s(1, 3) = s(1, 4) = ¼ and therefore s(i, j) = ¼, i, j = 1, 2, 3, 4.

It was noted in Tables 3 and 4 that for the one-parameter and two-parameter models, d(i, i, h) = s(i, i, 2h) at all epochs h, where s(i, i, 2h) is the probability that there is no substitution of base i at epoch 2h and d(i, i, h) is the probability that there is no difference between the two lineages for base i at epoch h. Consequently the complementary probabilities that there is a substitution and that there is a difference, respectively, are also equal. Note that this does not mean d(i, j, h) = s(i, j, 2h), i ≠ j, as was seen for the two-parameter model (Table 4).

This may also be proved more generally for any symmetric substitution matrix (Table 1, matrix VI) that is the same for both lineages. For such matrices s(i, j, h) = s(j, i, h). Then the probability that there are no substitutions at epoch 2h is $\sum_i$ s(i, i, 2h) = $\sum_i \sum_k s^2$(k, i, h). The probability that there is no difference between the two lineages at epoch h is $\sum_i$ d(i, i, h) = $\sum_{ik}$ s(k, i, h) s(k, i, h) = $\sum_{ik} s^2$(k, i, h). The argument does not hold for symmetric substitution matrices that are different for the two lineages, nor for asymmetric substitution matrices.

The equilibrium divergence matrix can be construed in two ways. First, the matrix may be construed as the probability that a base in the second lineage is T, C, A, or G conditioned on the probability that a base in the first lineage is T, C, A, or G, respectively, for each row. In this case the row sums will be 1; that is, the matrix is a conventional probability matrix (Table 5, row B). If the substitution matrices are symmetric, then in this case also if the composition of the first lineage is equal to the equilibrium conditional composition of the second lineage, the conditional composition of the second lineage will equal at all epochs its equilibrium value, and if the composition of the first lineage is not equal to the equilibrium conditional composition of the second lineage, the conditional composition of the second lineage will approach its equilibrium value with increasing epoch number k (Table 5, rows F and G). In the second construction each element d(i, j) is regarded as the probability that at the given site the base is j in the second lineage and i in the first lineage. In this case the sum of all elements in the matrix will be 1 (Table 5 row C). Note that if it is desired to fit an asymmetric observed divergence matrix, it is necessary to use asymmetric substitution matrices that are different for the two lineages (Table 5, row C).

Choice of the Number of Epochs for the Discrete Time Matrix Solution

The evolutionary distance between two lineages is the product of the number of epochs since divergence and the probability per epoch that a base is substituted. Table 6 shows that the graininess of the discrete time matrix solution does not introduce much uncertainty into this determination, since the evolutionary distance varies less than 2% for an eightfold increase in the number of epochs and a corresponding decrease in the probabilities of substitution. In the estimation of evolutionary trees the influence of even this small variability can be removed by fixing the number of epochs for all divergence matrices and using the estimated substitution rates.

Removal of an Ambiguity in the Assignment of Different Mutation Rates to Two Descendant Lineages

A more disconcerting property of the discrete matrix solution is that when one estimates the mutation rates in the two lineages separately, although the total mutation rate is well determined, the fraction of it assigned to each lineage is arbitrary, depending

Table 7. Determination of different substitution rates for two diverging lineages from observed divergence matrix

A	B	C	D	E	F	G	H	I
62,499	31,248	46,879	39,064	250,002	2000	1200	1000	64
60,608	27,991	48,966	42,158	249,872	2165	1161	875	64
53,791	20,392	55,929	49,038	248,580	2638	1141	614	64
74,444	41,757	33,299	28,419	248,095	1783	1172	1752	64

Columns A–E have been multiplied by 1,000,000; columns F–H by 1000; column I, by 1,000,000,000. A, Transition substitution rate of lineage 1 determined from best fit to observed divergence matrix. (Row 1 values are those used in generating the divergence matrix.) B, Same as A for ½× transversion rate. C, Same as A for lineage 2. D, Same as B for lineage 2. E, Total substitution rate [= A + C + 2(B + D), where letters denote values in corresponding columns]. F, A/B. G, C/D. H, (A + 2B)/(C + 2D). I, Sum of squares of residuals of fit of calculated best-fit divergence matrix to observed divergence matrix

on the initial values assumed in the iterative optimization process (Table 7). For a single pair of lineages there is no way to remove this ambiguity. However, if four or more lineages are available for the same gene, the information is sufficient to resolve the ambiguity using the evident topology of the evolutionary tree.

Let the evolutionary tree for four genes be as shown in Fig. 1. Let the true average transition substitution rates from a common ancestral time of -1 to the present time 0 be a, b, c, and d for lineages A, B, C, and D, respectively. Let the time of divergence of lineage B from the common ancestor of C and D be $-s$ and the time of divergence of C from D be $-t$. Let the rates of substitution in lineages A and B, estimated by optimization from the divergence matrix of A and B, be x_1 and x_2, respectively, and let the fraction of the total assigned to A be f_1. Similarly, for the pairs (A, C), (A, D), (B, C), (B, D), and (C, D), let the estimated rates be $x_3, x_4, x_5, \ldots, x_{12}$ and the fractions of the total rate assigned to the left members of the pairs be f_2, f_3, f_4, f_5, and f_6, respectively. Then the following relations hold:

$$f_1a = x_1 \quad (A1)$$
$$(1 - f_1)b = x_2 \quad (A2)$$
$$f_2a = x_3 \quad (A3)$$
$$(1 - f_2)c = x_4 \quad (A4)$$
$$f_3a = x_5 \quad (A5)$$
$$(1 - f_3)d = x_6 \quad (A6)$$
$$sf_4b = x_7 \quad (A7)$$
$$s(1 - f_4)c = x_8 \quad (A8)$$
$$sf_5b = x_9 \quad (A9)$$
$$s(1 - f_5)d = x_{10} \quad (A10)$$
$$tf_6c = x_{11} \quad (A11)$$
$$t(1 - f_6)d = x_{12} \quad (A12)$$

where a, b, c, d, s, t, f_1, f_2, f_3, f_4, f_5, and f_6 are to be determined from the calculated $x_1, x_2, x_3, \ldots, x_{12}$ estimated from the six observed divergence matrices by the optimization process. Equations (A1)–(A12) are 12 equations in 12 unknowns and may be solved as follows: From Eqs. (A1) and (A3) obtain

$$f_2 = f_1x_3/x_1 \quad (B1)$$

and similarly

$$f_3 = f_1x_5/x_1 \quad (B2)$$
$$sf_4 = (1 - f_1)x_7/x_2 \quad (B3)$$
$$sf_5 = (1 - f_1)x_9/x_2 \quad (B4)$$
$$tf_6 = (1 - f_2)x_{11}/x_4 = (1 - f_1x_3/x_1)x_{11}/x_4 \quad (B5)$$

Substitute Eq. (B1) in Eq. (A4) and obtain

$$(1 - f_1x_3/x_1)c = x_4 \quad (C1)$$

and similarly

$$(1 - f_1x_5/x_1)d = x_6 \quad (C2)$$
$$[s - (1 - f_1)x_7/x_2]c = x_8 \quad (C3)$$
$$[s - (1 - f_1)x_9/x_2]d = x_{10} \quad (C4)$$
$$[t - (1 - f_1x_3/x_1)x_{11}/x_4]d = x_{12} \quad (C5)$$

Substitute Eq. (C1) in Eq. (C3) and obtain

$$[s - (1 - f_1)x_7/x_2]x_4/(1 - f_1x_3/x_1) = x_8 \quad (D1)$$

and similarly

$$\frac{[s - (1 - f_1)x_9/x_2]x_6}{(1 - f_1x_5/x_1)} = x_{10} \quad (D2)$$

$$\frac{[t - (1 - f_1x_3/x_1)x_{11}/x_4]x_6}{(1 - f_1x_5/x_1)} = x_{12} \quad (D3)$$

Now Eqs. (D1) and (D2) are two linear equations in unknowns s and f_1 and are easily solved. Substitute f_1 into Eq. (D3) to obtain t; f_1 into Eqs.(A1), (A2), (C1), and (C2) to obtain a, b, c, and d; and f_1, s, and t into Eqs. (B1), (B2), (B3), (B4), and (B5) to obtain f_2, f_3, f_4, f_5, and f_6.

Then $k[f_1a + (1-f_1)b] = dAB$, the evolutionary distance between present-day lineages A and B corrected for multiple substitutions at any site in the sequence, where k is the epoch used for all the estimations $x_1, x_2, x_3, \ldots, x_{12}$. Similar calculations give dAC, dAD, dBC, dBD, and dCD. The edge lengths of the assumed tree can then be estimated by the conventional least-squares method (Fitch and Margoliash 1967).

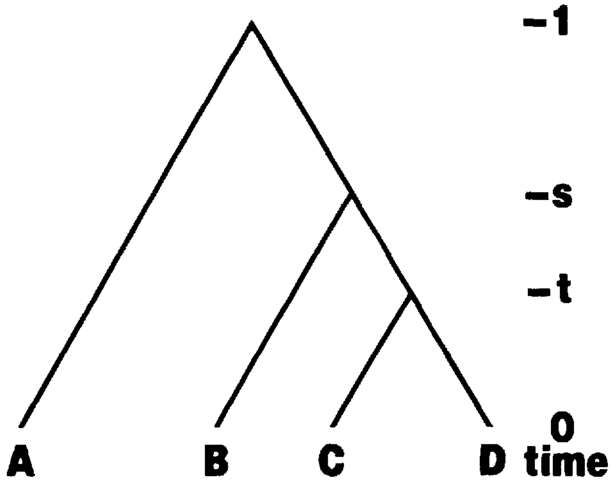

Fig. 1. Assumed evolutionary tree

Table 8. Comparison of observed rabbit–mouse beta-globin divergence matrix with results of four simulations of substitutions in two lineages by one-parameter and two-parameter models

	A	B
	One parameter	
	0	18
	3	20
	1	15
	2	16
Ave.	1.5	17.25
	Two parameter	
	7	13
	22	13
	18	24
	30	19
Ave.	19.25	17.25

A, Number of cases in 1000 trials in which T–C count was ≥11 and C–T count was ≤5. B, Number of cases in 1000 trials in which C–G count was zero and G–C count was ≥5

A similar solution can be obtained for each substitution-rate parameter set $x_1, x_2, x_3, \ldots, x_{12}$ estimated by the optimization procedure. In the two-parameter model (Table 1, matrix II) there are two such parameter sets, one for transitions and one for transversions. The evolutionary trees estimated from these two sets would be expected to be much the same, though it is obvious that after a few different substitutions have been established in each lineage the changed sequences may modify in different ways the probabilities of further substitutions, and that these modifications may be different for different classes of substitutions, for example, transitions and transversions. It is also possible to pool the evolutionary distances for the various substitution classes to obtain an overall or total evolutionary distance and to estimate the edge lengths of the assumed tree using the pooled data.

If data for the same gene exist for n lineages, where $n > 4$, the n overdetermined true average substitution rates a, b, c, . . . , n for the n lineages, the $n - 2$ relative branch times, and the n choose $2 = m$ fraction assignments $f_1, f_2, f_3, \ldots, f_m$ of the total average substitution rates to the respective numbers of the m possible pairs of lineages may be estimated from the 2m rates $x_1, x_2, x_3, \ldots, x_{2m}$ by nonlinear least-squares optimization. The x values themselves will have been estimated previously for each model parameter from the m divergence matrices by nonlinear least-squares optimization as described above. For example, if $n = 5$, then $n - 2 = 3$, $m = 10$, $2m = 20$, and there are $20 - 5 - 3 - 10 = 2$ degrees of freedom for the former estimation.

Choice of a Suitable Model

Since the speed of and avoidance of failure in the optimization process improve with the number of degrees of freedom in the estimation (15 minus the number of parameters estimated), it is desirable to keep the number of estimated parameters as small as will still permit fitting obvious features of the divergence matrix. For example, consider the beta-globin divergence matrix for rabbit (van Ooyen et al. 1979) and mouse (Konkel et al. 1979), matrix RM in Table 1. This matrix is for codon site 3 and for only those amino acids not paired with gaps in the alignment of Dayhoff (1978) for many alpha- and beta-globins. The meaning of the display has been described above. What model is suitable for these measurements? It appears that the transition values (11, 5, 4, and 4) are greater than the transversion values (1, 1, 1, 0, 3, 1, 5, and 5). A t-test for the difference between their means gives $t = 2.27$, $P = 0.025$. In fact, a t-test on all 36 possible pairs of 9 beta-like globins finds $t = 8.28$, $-10 \log P \approx 162$ (data not shown). Therefore it seems that a model of at least two parameters, one for transitions and one for transversions, is needed. The count of sites having C in mouse and T in rabbit, 11, is different from the count of sites having T in mouse and C in rabbit, 5, at aligned sites. Similarly, the count for C–G, 0, is different from that for G–C, 5. Table 8 gives the results of four Monte Carlo simulations of 1000 trials each for a one-parameter model in which the parameter equals the average, 3.416, and for a two-parameter model in which the transition parameter equals its average, 6, and the transversion parameter equals its average, 2.125. The values tabulated are the number of cases in which the T–C count is ≥11 and the C–T count is ≤5 and the number of cases in which the C–G count is zero and the G–C count is ≥5, respectively. For both T–C and C–G, the probability of such an asymmetry is <0.02, a clear indication that the model should provide an asymmetric divergence matrix. From the summary of solution properties in Table 5, such an asymmetric divergence matrix is attainable only from asymmetric substitution matrices that are different for the two lineages.

Application of the Discrete Time Matrix Method to an Observed Divergence Matrix

Discrete time matrix optimization solutions for the observed rabbit–mouse beta-globin divergence matrix (Table I, matrix RM) are given in Table 9 for five models with one to eight substitution-rate parameters plus three ancestral composition parameters. The five models are as follows: model (1), one parameter the same for both lineages (Table 1, matrix I); model (2), two parameters, symmetric, the same for both lineages (Table 1, matrix II); model (4), four parameters, asymmetric, the same for both lineages (Table 1, matrix VII); model (2,2), two parameters, symmetric, different for the two lineages (Table 1, matrix II); and model (4,4), four parameters asymmetric, different for the two lineages (Table 1, matrix VII).

The Euclidean norm of the residuals, rn, is largest for the one-parameter solution, (1); about the same for the two- and four-parameter solutions, (2), (4), and (2,2); and smallest for the eight-parameter solution, (4,4). The respective norms of the solution parameters, sn (Lawson and Hanson 1974), generally increase as the number of parameters increases. The substantial decrease in rn for model (4,4) is obtained at the cost of only a modest increase in sn, and I conclude that model (4,4) provides the best representation of the data. However, the best fit by the F-test is for model (2), for which $-10 \log P = 103$. S. Karlin (personal communication) has questioned this application of the F-test.

The estimated ancestral compositions, rows c1, c2, and c3 (and c4 by difference), are about the same for all five models. The relatively small fraction of base A is the most variable and for all models substantially higher than the observed values for mouse or rabbit (Table 1, matrix RM). For the best solution, model (4,4), the estimated fraction of T is about the same as that observed, the estimated fraction of C is intermediate between the two observed values, and the estimated fraction of G is higher than either of the observed values.

Table 9. Results of determination from observed rabbit–mouse beta-globin divergence matrix of ancestor composition and substitution rates in each of two lineages for three composition and one, two, four, or eight substitution-rate parameters

	Model				
	(1)	(2)	(4)	(2,2)	(4,4)
Xa	14,752	29,580	45,915	32,836	19,549
Xb		8,224	8,347	6,586	7,655
Xc			22,737		21,582
Xd			8,228		5
Ya				26,297	69,580
Yb				9,832	5,312
Yc					24,516
Yd					20,142
c1	253,349	251,931	290,141	250,652	279,179
c2	331,442	333,420	307,265	333,745	303,071
c3	16,227	10,461	18,827	11,041	20,006
c4	398,982	404,188	383,767	404,562	397,744
rn	69	49	46	48	15
sn	836	937	1,128	1,111	1,395
F	38	150	94	87	275
df	4,11	5,10	7,8	7,8	11,4
−10 log P	56	103	62	63	45

All values in rows Xa to sn have been multiplied by 1,000,000. Codes for rows: X, first lineage; Y, second lineage; c, base compositions; a, b, c, d, parameter designations (Table I, matrices I, II, VII); rn, square root of sum of squares of residuals; sn, square root of sum of squares of estimated parameters; F, F-value for fit to observed divergence matrix; df, degrees of freedom. Model (1), one parameter the same for both lineages. Model (2), two parameters the same for both lineages. Model (4), four parameters the same for both lineages. Model (2,2), two parameters for each of the two lineages. Model (4,4), four parameters for each of the two lineages

The substitution-rate parameters appear to vary widely for the five models, but closer inspection finds them to be relatively consistent. For example, one-third the transition rate plus two-thirds the transversion rate for model (2) approximately equals the total rate for model (1): ⅓ (29,580) + ⅔ (8224) = 15,343 ≈ 14,752. A weighted average of the rates a and b for the two lineages in model (2,2) equals the common rate of the two lineages assumed to be the same in model (2): 0.502(32,836) + 0.498(26,297) = 29,580 and 0.495(6586) + 0.505(9832) = 8224. Similarly, a weighted average of the rates of transition substitution from T to C and C to T and from A to G and G to A in model (4) equals the common rate for both directions in model (2): 0.295(45,915) + 0.705(22,737) = 29,580. The forward, backward, and combined transversion rates are all about the same: 8347 ≈ 8228 ≈ 8224. A weighted average of three of the four rates for model (4,4) equals the rate for model (4):

$$0.473(19{,}549) + 0.527(69{,}580) = 45{,}915$$
$$0.606(21{,}582) + 0.394(24{,}516) = 22{,}737$$
$$0.592(5) + 0.408(20{,}142) = 8228$$

Only for rates xb/yb does such an average fail, since both xb/yb for model (4,4) are less than xb for model (4), a divergence that may explain in part the much better fit for model (4,4), in which the residual norm, rn, is reduced by a factor of ⅓.

Table 10. Calculated divergence matrices

Model		T	C	A	G
RM	T	25	11	1	1
	C	5	33	1	0
	A	3	1	2	4
	G	5	5	4	40
(1)	T	25.4	3.9	2.0	4.3
	C	3.9	32.1	2.5	4.8
	A	2.0	2.5	2.0	2.9
	G	4.3	4.8	2.9	39.8
(2)	T	25.3	7.2	1.3	2.6
	C	7.2	33.1	1.6	2.8
	A	1.3	1.6	1.7	5.1
	G	2.6	2.8	5.1	39.7
(4)	T	25.1	8.0	1.3	2.6
	C	8.0	33.0	1.5	3.0
	A	1.3	1.5	2.0	4.1
	G	2.6	3.0	4.1	40.0
(2,2)	T	25.1	7.3	1.5	2.5
	C	7.1	33.2	1.8	2.8
	A	1.1	1.4	1.8	5.6
	G	2.6	2.8	4.6	39.8
(4,4)	T	25.1	10.9	0.7	0.8
	C	5.1	33.0	0.7	0.8
	A	1.5	2.0	2.1	4.1
	G	4.6	5.5	4.1	40.0

The divergence matrices calculated from the estimated parameters of the five models are displayed in Table 10. The calculated matrix for model (1) is symmetric and the average difference between the two sequences is 3.4, in good agreement with the observed average of 3.416. The calculated matrix for model (2) is also symmetric, with an average transition difference of 6.15 and transversion difference of 2.1, in good agreement with the observed values of 6.0 and 2.125. The solution also provides a larger transition difference for the (T,C) pair than for the (A,G) pair, in agreement with observation but not with the correct average values. It also fails to provide a larger average for the lower-left-hand 2 × 2 transversion submatrix than for the upper-right-hand submatrix, as is seen in the observed matrix. The calculated matrix for model (4) is also symmetric, and it provides the correct average values for the (T,C) and (A,G) pair transitions. It again fails to provide a larger average for the lower-left-hand transversion submatrix than for the upper-right-hand one. The calculated matrix for model (2,2) is much like that for model (2). The calculated matrices for models (2), (4), and (2,2) all give larger values for G–T and G–C than for A–T and A–C, in agreement with observation, but because of their symmetry [or near symmetry for model (2,2)] introduce a difference where none is warranted in the upper-right-hand submatrix. The calculated matrix for model (4,4) exhibits most of the features of the observed matrix and is in close agreement with it. There are more Cs in mouse aligned with Ts in rabbit, 10.9, than there are Ts in rabbit aligned with Cs in mouse, 5.1, but the numbers of Gs in mouse aligned with As in rabbit and of As in mouse aligned with Gs in rabbit are the same, 4.1. Note that the latter value is less than the former two. The G–T and G–C differences are larger than the A–T and A–C differences, and the averages of both, 5.05 and 1.75, respectively, are close to the observed values of 5.00 and 2.00. The T–G and C–G differences are about the same as the T–A and C–A differences, and their overall average, 0.75, is equal to the observed value. From this examination of the details of the calculated fit, it is concluded that model (4,4) provides the best representation of the data, in agreement with the conclusion reached above from the trend in the residual norm with the solution norm.

Acknowledgments. I am grateful to E. Zuckerkandl for asking me to examine the statistical properties of DNA sequences and for supporting me in this pursuit. I am also grateful to S. Karlin for helpful comments as reviewer. I thank S. Schwoebel for typing the manuscript.

References

Box MJ (1965) A new method of constrained optimization and a comparison with other methods. Comput J 8:42–52

Dayhoff MO (ed) (1978) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, DC

Feller W (1968) An introduction to probability theory and its applications, vol 1, 3rd ed. John Wiley and Sons, New York

Fitch WM (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta homoglobulin messenger RNA's. J Mol Evol 16:153–209.

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Gojobori T, Ishii K, Nei M (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J Mol Evol 18:414–422

Holmquist R (1972) Theoretical foundations for a quantitative approach to paleogenetics. J Mol Evol 1:115–133

Holmquist R (1976) Solution to a gene divergence problem under arbitrary stable nucleotide transition probabilities. J Mol Evol 8:337–349

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro NH (ed) Mammalian protein metabolism, Academic Press, New York, pp 21–123

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Estimation of evolutionary differences between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Konkel DA, Maizel JV, Leder P (1979) The evolution and sequence comparison of two recently diverged chromosome beta globin genes. Cell 18:865–873

Lanave C, Preparata G, Saccone C, Semo G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs, New Jersey

Pauling L, Zuckerkandl E (1963) Chemical paleogenetics, molecular "restoration studies" of extinct forms of life. Acta Chem Scand 17(suppl 1):59–516

Powell MJD (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives. Comput J 7:155–162

Spendley W, Hext GR, Himsworth FR (1962) Sequential applications of simplex designs in optimization and evolutionary operation. Technometrics 4:441–461

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269–285

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudo-genes. Genetics 98:641–657

van Ooyen A, VandenBerg J, Mantei N, Weissman, C (1979) Comparisons of total sequence of a cloned rabbit beta globin gene and its flanking regions with a homologous mouse sequence. Science 206:337–344

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New York, pp 189–225

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166

Received December 4, 1984/Revised March 15, 1985