J. theor. Biol. (1990) **142,** 485-501

The General Stochastic Model of Nucleotide Substitution

F. RODRÍGUEZ, J. L. OLIVER†, A. MARÍN AND J. R. MEDINA‡†

Departamento de Genética, Facultad de Biología, Universidad de Sevilla, Ap. 1095, *E*-41080-*Sevilla, Spain, and* † *Unidad de Genética, Facultad de Ciencias, Universidad de Granada, E*-18071-*Granada, Spain*

(*Received on* 18 *May* 1989, *Accepted in revised form* 29 *September* 1989)

DNA sequence evolution through nucleotide substitution may be assimilated to a stationary Markov process. The fundamental equations of the general model, with 12 independent substitution parameters, are used to obtain a formula which corrects the effect of multiple and parallel substitutions on the measure of evolutionary divergence between two homologous sequences. We show that only reversible models, with six independent parameters, allow the calculation of the substitution rates. Simulation experiments on DNA sequence evolution through nucleotide substitution call into question the effectiveness of the general model (and of any other more detailed description); nevertheless, the general model results are slightly superior to any of its particular cases.

Introduction

In the study of molecular evolution, DNA sequences are more informative than protein sequences, since there are synonymous codons as well as many DNA sequences which are not translated to protein. The evolutionary change of DNA occurs either by nucleotide substitution, or by various deletion and insertion processes. We address here the first of these processes, that is estimated from the average number of nucleotide substitutions per nucleotide site between two homologous DNA sequences.

When the two sequences diverged a long time ago, multiple and parallel changes may have taken place; if so, the observed proportion of different nucleotides between the two sequences must be an underestimate of the number of nucleotide substitutions. To avoid this underestimation, various statistical methods have been proposed. All these methods rely on models of nucleotide substitution that are particular cases of the general 4 hypothesis model (G4H).

The G4H Model

The G4H model assumes that nucleotide substitution is a Markov process in which the rates of substitution are: (a) not dependent on the sequence site; (b) constant in time; (c) the same for the two lineages; (d) base frequencies of the ancestral sequence are equilibrium frequencies, so that they remain unaltered during the process. Note that these four hypotheses are mutually independent; in particular,

‡ Author to whom all correspondence should be addressed.

0022-5193/90/040485+17 $03.00/0

equilibrium frequencies are dependent on the relative values of the different rates of substitution, but not on their absolute values.

Parameters of the G4H Model

The general model (G4H), which has 12 independent substitution rates and only assumes the four mentioned hypotheses, uses the following 4×4 matrices:

(1) The divergence matrix $X(t) = [x_{ij}(t)]$, where $x_{ij}(t)$ is the probability that, at time t, at a given nucleotide site, the first sequence has base i and the second sequence has base j. From hypothesis d it follows that, for $i \neq j$, $x(0)_{ij} = 0$, and for $i = j$, $x(0)_{ij} = q_i$, where q_i is the equilibrium frequency of base i. That is, the divergence matrix is at the beginning a diagonal matrix whose elements of the main diagonal are the equilibrium frequencies of the ancestral sequence.

(2) The matrix of substitution rates, $R = (r_{ij})$, where the elements of R are the transient intensity functions of the stochastic process; that is, for $i \neq j$, r_{ij} is the rate of substitution for base i by the base j, and for $i = j$, $r_{ii} = -\sum_{j \neq i} r_{ij}$. Note that the matrix of substitution rates has a maximum of 12 independent elements.

(3) The evolutionary matrix, $P(t) = [p_{ij}(t)]$, where $p_{ij}(t)$ is the conditional probability that, at a given nucleotide site, there is base j at time t, given base i at initial time. Of course, since the two sequences are the same at the beginning, $P(0) = I$, where I is the identity matrix.

(4) For computer simulation, time must be considered as a discrete variable; therefore, instead of the matrix R, we use as a *transition matrix* between two consecutive generations the matrix $S = (s_{ij})$, where s_{ij} is the probability that base i in generation g is substituted by base j in generation $g+1$. Although

$$s_{ii} = 1 - \sum_{j \neq i} s_{ij},$$

the form of S is equal to the form of R for each substitution scheme.

Particular Cases of the G4H Model

Every specific model previously worked out introduces some additional hypotheses which simplifies the G4H model (Table 1).

The JC (Jukes & Cantor, 1969), the K2P (Kimura, 1980), the K3ST (Kimura, 1981) and the L (Lanave *et al.*, 1984) models share the property that the transition matrix of substitution rates is symmetric; each of these models can be reduced to the preceding one by imposing more restrictive conditions; thus each one is a particular case of the next.

The models TK (Takahata & Kimura, 1981), GIN (Gojobori *et al.*, 1982) and TN (Tajima & Nei, 1984) do not have symmetric matrices, and have five, six and four parameters, respectively.

All these previous models, including the L model, are particular cases of G4H model.

TABLE 1

Transition matrices of the different models of nucleotide substitution. For symmetric models only the pertinent hemimatrix is showed. The JC model assumes that all kinds of base substitutions are equally likely, so that the rate of substitution is the same for all nucleotide pairs. In the K2P model, a difference between the rate of transition type substitutions and that of transversion type substitutions is considered, so that a two parameter model results. In the K3ST model, three different rates of substitution are considered: one for transitions, the other for transversions between two complementary bases (e.g. A by T) and a third for transversions between no complementary bases (e.g. A by C). In the L model, six parameters are considered, conforming a symmetric matrix of substitution rates, in which each element of the hemimatrix has a different value. In the TN model, the rate of substitution of nucleotide j for nucleotide i per unit evolutionary time is the same irrespective of nucleotide i. The models TK, GIN and B do not have symmetric matrices, and have five, six and four parameters, respectively

JC

		A	T	C	G
K2P	A	—	α	α	α
	T	β	—	α	α
	C	β	α	—	α
	G	α	β	β	—

K3ST

		A	T	C	G
L	A	—	β	γ	α
	T	α	—	α	γ
	C	β	δ	—	β
	G	κ	ε	ϕ	—

TN

		A	T	C	G
	A	—	β	γ	ε
	T	α	—	γ	ε
	C	α	β	—	ε
	G	α	β	γ	—

TK

		A	T	C	G
	A	—	γ	δ	α
	T	γ	—	α	δ
	C	ε	β	—	γ
	G	β	ε	γ	—

GIN

		A	T	C	G
	A	—	$\alpha 1$	α	α
	T	$\beta 1$	—	α	α
	C	β	β	—	$\alpha 2$
	G	β	β	$\beta 2$	—

B

		A	T	C	G
	A	—	α	β	β
	T	κ	—	β	β
	C	δ	δ	—	α
	G	δ	δ	κ	—

G4H

		A	T	C	G
	A	—	α	β	χ
	T	δ	—	ε	ϕ
	C	γ	η	—	ι
	G	ψ	κ	λ	—

The B model (Blaisdell, 1985), which takes into account four parameters, is not symmetric, and, at variance with the methods mentioned above, it does not require either that the nucleotide frequencies of the ancestral sequence are at equilibrium, or that the substitution rates in the two descendant sequences are the same. The Blaisdell model allows the estimation of the substitution rates by minimizing the sum of squares of the differences between the predicted and the observed divergence matrices. It is, therefore, a semiempiric method which cannot be reduced to the G4H model.

Fundamental Equations of the G4H Model

With the above notation, the fundamental equations of the G4H model are:

$$X(t) = P'(t)X(0)P(t) \quad \text{or} \quad x(t)_{ij} = \sum_{k=1}^{k=4} p'(t)_{ik} q_k p(t)_{kj}, \tag{1}$$

$$dP(t)/dt = P(t)R \quad \text{or} \quad dp(t)_{ij}/dt = \sum_{k=1}^{k=4} p(t)_{ik} r_{kj}, \tag{2}$$

$$P(t) = \exp(Rt) \quad \text{or} \quad P(t) = I + Rt + \ldots + R^n t^n / n! + \ldots, \tag{3}$$

where $p'(t) = [p'(t)_{ij}]$ is the transpose of $p(t)$, that is,

$$p'(t)_{ij} = p(t)_{ji}.$$

From eqn (1) it follows that $X(t)$ is a symmetric matrix, that is, $x(t)_{ij} = x(t)_{ji}$.

So, eqn (1) is not a system of 15 independent linear equations; given the symmetry of the $X(t)$ matrix and the fact that

$$2x(t)_{ii} = q_i - \sum_{k \neq i} x(t)_{ik} - \sum_{k \neq i} x(t)_{ki},$$

it turns out that eqn (1) has only six independent linear equations. As a consequence, the substitution rates may be obtained only in models having at most six independent parameters.

The observed divergence matrix is usually asymmetric; this may be due either to the fact that the actual evolutionary process does not follow the conditions imposed by the model, or to the sampling error derived from the finite length of the sequences used.

Because the observed divergence matrix is asymmetric, all the formulae to correct the effect of multiple and parallel changes, including that of the G4H model, may become inapplicable to a given divergence matrix, because one of the arguments of the logarithms or square roots involved becomes negative (Kimura, 1981). The proportion of inapplicable cases constitutes a measure of the robustness of a model.

Expected Number of Substitutions

Although the values of the substitution rates cannot be, in general, obtained, for many evolutionary studies it suffices to determine the expected number of nucleotide substitutions per site, given by $\delta = 2kt$; the value of δ may be obtained under any

of the described specific models. Here we report a more general way of estimating δ. See Appendix I for a detailed rationale,

$$\delta = -\text{tr}\,[(X(0))\log(X(0)^{-1}X(t))]. \tag{4}$$

This expression for the average number of nucleotide substitution is valid for any scheme that satisfies

$$\text{tr}\,[X(0)F] = 0,$$

where $F[X(0)^{-1}R'tX(0), Rt]$ is a polynomial in its matrix arguments whose trace is 0.

Although this formula does not reduce to the GIN formula, our formula is valid under a wider class of substitution schemes. Our formula can be computed without using approximate or iterative methods.

Particular Cases

Two families of models that satisfy the above condition are interesting:

(a) The commutative models, in which the arguments of F commute, i.e.

$$X(0)^{-1}R'tX(0)Rt = RtX(0)^{-1}R'tX(0).$$

See Appendix II for an explicit proof that $\text{tr}\,(F) = 0$.

An interesting subclass of the commutative family are the reversible models, in which both arguments of F are equal:

$$X(0)^{-1}R'tX(0) = Rt,$$

i.e.

$$X(0)^{-1}R'X(0) = R$$

or

$$R'X(0) = X(0)R. \tag{5}$$

Equation (5) implies that R has a maximum of six independent parameters; the remaining six parameters would depend on the first six rates and the equilibrium frequencies. For example, the TK model has four independent parameters and $q_1 = q_3$ and $q_2 = q_4$.

The L model is the most general reversible model. While in the G4H model the 12 substitution rates cannot be estimated without additional assumptions, since, as previously explained, there are less independent linear equations than parameters to obtain in eqn (1), the estimation of at most six independent parameters of the reversible models is possible, as they satisfy

$$\log[X(0)^{-1}X(t)] = 2Rt.$$

We call them reversible because from eqn (5) it follows, as explained in Appendix III, that

$$X(0)P(t) = P'(t)X(0)$$

or

$$q_i p(t)_{ij} = q_j p(t)_{ji}. \tag{6}$$

In evolutionary terms, eqn (6) means that if we compare any two sequences with equilibrium nucleotide compositions, the probability of evolving from the first to the second one is equal to the probability of evolving from the second to the first: that is, evolution is freely reversible along every equilibrium path. This evolutionary reversibility implies that, given a present sequence and an ancestral sequence, we should not be able to tell which is the present and which the ancestral one. In fact, given this equivalence of present and past along every simple evolutionary path, we need at least two homologous sequences at two times to know the sense of "the arrow of time"; the time flows from the state in which the two sequences are less different to the state in which they are relatively more different.

(b) Equifrequency models, which lead to equal equilibrium frequencies of the four nucleotides, so that

$$\mathrm{tr}\,[X(0)F] = \mathrm{tr}\,[(0{\cdot}25I)F] = 0{\cdot}25\,\mathrm{tr}\,(IF) = 0{\cdot}25\,\mathrm{tr}\,(F) = 0.$$

Any scheme with a symmetric R matrix is an equifrequency model. It is also a reversible model: $R = R'$ implies $X(0)R = R'X(0)$, as $X(0)$ is a diagonal matrix with $\forall_{i,j} x(0)_{ii} = x(0)_{jj}$. Such are the JC, K2P, K3ST and L models.

More generally: any model in which $(I+R)$ is a double stochastic matrix, that is

$$\sum_{j=1}^{j=4} r_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{i=4} r_{ij} = 0,$$

for i or j equal to 1, 2, 3 and 4, is an equifrequency model. These extra relations between the substitution rates allow for a maximum of eight independent parameters. All the equifrequency models with asymmetric R are not reversible, since $r_{ij} \neq r_{ji}$ implies that $X(0)R \neq R'X(0)$ when $\forall_{i,j} x(0)_{ii} = x(0)_{jj}$.

A Numerical Example

As an example, we have applied formula (6) to compare the third codon positions of the nucleotide sequences of the β-globin genes from mouse (Konkel *et al.*, 1979) and rabbit (Hardison *et al.*, 1979). The results obtained are presented in Table 2:

(1) This matrix represents the observed divergence between the two sequences. It gives the 16 combinations of base pairs when the homologous sites in the two sequences are compared. The four elements in the first row represent, respectively, the cases in which sites in the rabbit sequence are occupied by bases A, T, C and G, while A occurs in the mouse sequence; the second row corresponds to sites occupied by bases A, T, C and G in the rabbit sequence, occurring T in the mouse sequence, and similarly for the other two rows.

(2) This matrix $X(t)$ contains the relative nucleotide pair frequencies, which are obtained by dividing the corresponding elements of the observed divergence matrix by 146, which is the number of codons in the β-globin gene.

(3) This matrix $X(0)$ is a diagonal matrix containing the equilibrium frequencies of bases A, T, C and G, these have been estimated as the average values for the two sequences in the observed divergence matrix.

(4) This matrix $X(0)^{-1}$ is the inverse of $X(0)$.

TABLE 2

Numerical example: estimates of the number of nucleotide substitutions per site between the mouse and rabbit β-globin genes at the third codon position

(1)		A	T	C	G
	A	1	1	1	5
	T	3	26	6	5
	C	1	14	33	5
	G	4	1	0	40
(2)	0·007	0·007	0·007	0·034	
	0·021	0·178	0·041	0·034	
	0·007	0·096	0·226	0·034	
	0·027	0·007	0·000	0·274	
(3)	0·058	0·000	0·000	0·000	
	0·000	0·281	0·000	0·000	
	0·000	0·000	0·318	0·000	
	0·000	0·000	0·000	0·343	
(4)	17·241	0·000	0·000	0·000	
	0·000	3·559	0·000	0·000	
	0·000	0·000	3·145	0·000	
	0·000	0·000	0·000	2·915	
(5)	0·121	0·121	0·121	0·586	
	0·075	0·633	0·146	0·121	
	0·022	0·302	0·711	0·107	
	0·079	0·020	0·000	0·799	
(6)	0·051	0·958	0·453	0·801	
(7)	0·988	0·363	0·077	−0·182	
	−0·115	0·529	−0·682	0·629	
	0·036	0·786	0·784	1·328	
	−0·101	0·247	0·023	−0.615	
(8)	0·936	−0·091	−0·151	−0·698	
	0·185	0·529	0·403	1·357	
	−0·089	−0·858	0·528	0·289	
	−0·082	0·196	0·207	−0·955	
(9)	−2·972	0·000	0·000	0·000	
	0·000	−0·043	0·000	0·000	
	0·000	0·000	−0·791	0·000	
	0·000	0·000	0·000	−0·221	
(10)	−2·745	0·320	0·413	1·970	
	0·278	−0·534	0·196	0·020	
	−0·026	0·467	−0·386	0·130	
	0·269	0·009	−0·031	−0·359	
(11)	−0·159	0·018	0·024	0·114	
	0·078	−0·150	0·055	0·006	
	−0·008	0·148	−0·123	0·041	
	0·092	0·003	−0·011	−0·123	
(12)	G4H = 0·555		K2P = 0·426		
	JC = 0·408		TK = 0·463		
	K3ST = 0·426		TN = 0·465		
	GIN = 0·567				

(5) This is the product matrix $X(0)^{-1}X(t)$. Let us call Π this matrix. In order to compute the expression

$$\ln [X(0)^{-1}X(t)] = \ln \Pi$$

we use the following formula:

$$\ln \Pi = \Omega \Psi \Omega^{-1},$$

where Ω which is given in (7) is a matrix containing as columns the eigenvectors of Π, Ω^{-1} is the inverse of the above and is given in (8), Ψ given in (9) is a diagonal matrix whose elements are the natural logarithms of the eigenvalues of Π which are given in (6). The matrix $\ln \Pi$ computed in this way is given in (10).

Finally, the product matrix $X(0) \ln [X(0)^{-1}X(t)]$ is given in (11). The negative trace of the above matrix is the expected number of nucleotide substitutions and is given in (12) as G4H. Item (12) also shows other estimates of the number of nucleotide substitutions such as JC, K2P, K3ST, TK, GIN, and TN.

Simulation of the Nucleotide Substitution Process

In order to check the validity of our formula and determine its range of applicability in comparison with other models, we have performed a computer simulation of DNA evolution on finite sequences through strict stochastic process, although conforming to the particular substitution rates imposed by each scheme. The SDSE software package (Oliver *et al.*, 1989) was used for this purpose. The substitution schemes used were those of Jukes & Cantor (1969) single parameter (JC) method, Kimura's (1980) two-parameter (K2P) method, Kimura's (1981) three-substitution-type (K3ST) method, Takahata & Kimura's (1981) (TK) method, Gojobori *et al.*'s (1982) (GIN) method, and Tajima & Nei's (1984) (TN) method. In addition, we employed three schemes (RAND0, RAND1, and RAND2) generated by using random numbers as substitution rates, as well as those used by Gojobori *et al.* (1982).

For each scheme, the substitution matrix (R_{ij}, $i \neq j$) was converted into the transition matrix (S_{ij}) corresponding to

$$k \equiv \sum_i q_i \sum_{j \neq i} s_{ij} = 0{\cdot}01,$$

where k is the average number of nucleotide substitutions per site per unit evolutionary time. These matrices are given in Table 3. For convenience we used a discrete time, rather than a continuous time, approach. Equilibrium frequencies were computed by squaring the corresponding transition matrix repeatedly. Squaring was continued until all the elements in each column have the same value; the equilibrium composition is equal to any row and, as the process starts at the equilibrium composition, it stays at it.

For each scheme, we generated a random DNA sequence of N nucleotides at equilibrium frequencies, which was then used as an ancestral sequence in the

TABLE 3

Values for parameters of the different substitution schemes used in computer simulation

JC			
$\alpha = 0{\cdot}00333$			
$\hat{q}(A) = 0{\cdot}25$	$\hat{q}(T) = 0{\cdot}25$	$\hat{q}(C) = 0{\cdot}25$	$\hat{q}(G) = 0{\cdot}25$
K2P			
$\alpha = 0{\cdot}00111$	$\beta = 0{\cdot}00444$		
$\hat{q}(A) = 0{\cdot}25$	$\hat{q}(T) = 0{\cdot}25$	$\hat{q}(C) = 0{\cdot}25$	$\hat{q}(G) = 0{\cdot}25$
K3ST			
$\alpha = 0{\cdot}00444$	$\beta = 0{\cdot}00222$	$\gamma = 0{\cdot}00333$	
$\hat{q}(A) = 0{\cdot}25$	$\hat{q}(T) = 0{\cdot}25$	$\hat{q}(C) = 0{\cdot}25$	$\hat{q}(G) = 0{\cdot}25$
TK			
$\alpha = 0{\cdot}00152$	$\beta = 0{\cdot}00228$	$\gamma = 0{\cdot}00342$	
$\delta = 0{\cdot}00304$	$\varepsilon = 0{\cdot}00456$	$\theta = 2$	
$\hat{q}(A) = 0{\cdot}30$	$\hat{q}(T) = 0{\cdot}30$	$\hat{q}(C) = 0{\cdot}20$	$\hat{q}(G) = 0{\cdot}20$
GIN			
$\alpha = 0{\cdot}00237$	$\alpha_1 = 0{\cdot}00472$	$\alpha_2 = 0{\cdot}00709$	
$\beta = 0{\cdot}00118$	$\beta_1 = 0{\cdot}00591$	$\beta_2 = 0{\cdot}00827$	
$\hat{q}(A) = 0{\cdot}180$	$\hat{q}(T) = 0{\cdot}154$	$\hat{q}(C) = 0{\cdot}355$	$\hat{q}(G) = 0{\cdot}311$
TN			
$\alpha = 0{\cdot}00132$	$\beta = 0{\cdot}00264$	$\gamma = 0{\cdot}00396$	$\varepsilon = 0{\cdot}00528$
$\hat{q}(A) = 0{\cdot}1$	$\hat{q}(T) = 0{\cdot}2$	$\hat{q}(C) = 0{\cdot}3$	$\hat{q}(G) = 0{\cdot}4$
RAND0			
$\alpha = 0{\cdot}00119$	$\beta = 0{\cdot}00239$	$\chi = 0{\cdot}00478$	
$\delta = 0{\cdot}00478$	$\varepsilon = 0{\cdot}00358$	$\phi = 0{\cdot}00239$	
$\gamma = 0{\cdot}00358$	$\eta = 0{\cdot}00239$	$\iota = 0{\cdot}00478$	
$\psi = 0{\cdot}00358$	$\kappa = 0{\cdot}00119$	$\lambda = 0{\cdot}00597$	
$\hat{q}(A) = 0{\cdot}313$	$\hat{q}(T) = 0{\cdot}127$	$\hat{q}(C) = 0{\cdot}272$	$\hat{q}(G) = 0{\cdot}288$
RAND1			
$\alpha = 0{\cdot}0057$	$\beta = 0{\cdot}0070$	$\chi = 0{\cdot}0029$	
$\delta = 0{\cdot}0034$	$\varepsilon = 0{\cdot}0038$	$\phi = 0{\cdot}0052$	
$\gamma = 0{\cdot}0013$	$\eta = 0{\cdot}0032$	$\iota = 0{\cdot}0007$	
$\psi = 0{\cdot}0059$	$\kappa = 0{\cdot}0073$	$\lambda = 0{\cdot}0022$	
$\hat{q}(A) = 0{\cdot}145$	$\hat{q}(T) = 0{\cdot}263$	$\hat{q}(C) = 0{\cdot}454$	$\hat{q}(G) = 0{\cdot}138$
RAND2			
$\alpha = 0{\cdot}00581$	$\beta = 0{\cdot}00454$	$\chi = 0{\cdot}00018$	
$\delta = 0{\cdot}00654$	$\varepsilon = 0{\cdot}00063$	$\phi = 0{\cdot}00220$	
$\gamma = 0{\cdot}00492$	$\eta = 0{\cdot}00079$	$\iota = 0{\cdot}00392$	
$\psi = 0{\cdot}00718$	$\kappa = 0{\cdot}00065$	$\lambda = 0{\cdot}00248$	
$\hat{q}(A) = 0{\cdot}367$	$\hat{q}(T) = 0{\cdot}257$	$\hat{q}(C) = 0{\cdot}228$	$\hat{q}(G) = 0{\cdot}148$
GIN'82			
$\alpha = 0{\cdot}00125$	$\alpha_1 = 0{\cdot}008$	$\alpha_2 = 0{\cdot}0118$	
$\beta = 0{\cdot}005$	$\beta_1 = 0{\cdot}004$	$\beta_2 = 0{\cdot}0059$	
$\hat{q}(A) = 0{\cdot}290$	$\hat{q}(T) = 0{\cdot}510$	$\hat{q}(C) = 0{\cdot}079$	$\hat{q}(G) = 0{\cdot}121$

simulation. From this sequence, two descendent sequences were obtained by nucleotide substitution according to the schemes mentioned above.

We simulated the evolution of the ancestral sequence as follows. A nucleotide site in the sequence is chosen at random. Any of the four nucleotides (A, T, C or G) may be at this site. Let T, for example, be the nucleotide occupying this site. We divided the range 0–1 into four segments and used the transition matrix to assign the length of each segment; the length of the first segment is proportional to the probability that T does not change (T→T); the length of the three remaining segments were made proportional to the probabilities that T is substituted by A (T→A), by C (T→C) or by G (T→G). We then generated a random number in the range 0–1. The original T was not changed if this number lay within the first segment, but was substituted by A, C or G if the random number lay within the second, third or fourth segment, respectively. The same was done with any nucleotide occupying the chosen site. A similar process was repeated the desired number of steps, thus generating an "evolved" sequence. We then compared every pair of "evolved" sequences obtained and computed the matrix of nucleotide pair frequencies (x_{ij}) for each sequence pair. We applied the different methods of estimating nucleotide substitutions to these nucleotide pair frequency data to estimate evolutionary distances.

The proportion of inapplicable cases increases when the number of substitutions is large and the number of nucleotide sites compared is small. The effects of both on the estimates of δ and the number of cases in which the methods of estimation are inapplicable have been examined by Gojobori *et al.* (1982) and Tajima & Nei (1984). For computational convenience as well as to reduce the number of inapplicable cases, we have chosen sequences of 640 nucleotides in length. Simulation was allowed to proceed until the accumulated number of substitution steps between descendent sequences were $\delta = 0{\cdot}125$, $\delta = 0{\cdot}25$, $\delta = 0{\cdot}5$, $\delta = 1{\cdot}0$ and $\delta = 2{\cdot}0$. The simulation was repeated ten times for each scheme and level of divergence in the first six cases of Table 3, whereas it was repeated 50 times for the RAND1, RAND2 and GIN'82 schemes.

Results of computer simulation are presented in Table 4 for the substitution schemes corresponding to JC, K2P, K3ST, TK, GIN and TN methods, and in Table 5 for schemes RAND0, RAND1 and RAND2. The means and standard deviations of the estimates of δ have been obtained by excluding inapplicable cases when they occur; the number of such inapplicable cases are given in brackets close to the method of estimation involved.

Extensive comparisons among the estimates of δ obtained by different methods have been carried out by other authors. Takahata & Kimura (1981) compared TK, K3ST and JC formulae, their results showed that TK gives good estimates for $\delta < 1$ being more accurate than K3ST and JC. Gojobori *et al.* (1982) carried out computer simulation experiments under a variety of simulation schemes, comparing GIN, TK and JC methods; they concluded that although GIN formula is often inapplicable it gives a better estimate than the other two methods when it is applicable. Tajima & Nei (1984) and Tajima (1985) compared estimation methods 1 to 6, concluding that TN formula gives good estimates for a wide range of nucleotide substitution

TABLE 4

Means ($\bar{\delta}$) and S.Ds (σ_δ) of δ over ten replications. The number of inapplicable cases is given in brackets

True distance	Method	Substitution schemes: JC $\bar{\delta}$	JC σ_δ	K2P $\bar{\delta}$	K2P σ_δ	K3ST $\bar{\delta}$	K3ST σ_δ
0·125	JC	0·127	0·007	0·125	0·006	0·126	0·005
	K2P	0·128	0·006	0·127	0·006	0·127	0·006
	K3ST	0·128	0·006	0·127	0·006	0·128	0·006
	TK	0·127	0·007	0·126	0·006	0·127	0·006
	GIN	0·128	0·007	0·126	0·006	0·127	0·006
	TN	0·127	0·007	0·126	0·006	0·127	0·006
	G4H	0·127	0·006	0·126	0·006	0·127	0·006
0·25	JC	0·249	0·008	0·254	0·011	0·251	0·011
	K2P	0·249	0·007	0·257	0·012	0·252	0·011
	K3ST	0·249	0·007	0·257	0·012	0·252	0·011
	TK	0·250	0·008	0·257	0·011	0·253	0·011
	GIN	0·250	0·008	0·255	0·011	0·253	0·011
	TN	0·250	0·008	0·259	0·011	0·254	0·011
	G4H	0·250	0·008	0·256	0·010	0·253	0·011
0·50	JC	0·508	0·024	0·484	0·054	0·481	0·060
	K2P	0·509	0·025	0·493	0·057	0·485	0·062
	K3ST	0·509	0·025	0·493	0·057	0·486	0·063
	TK	0·509	0·024	0·493	0·057	0·486	0·063
	GIN	0·510	0·024	0·488	0·055	0·486	0·062
	TN	0·510	0·024	0·494	0·058	0·487	0·063
	G4H	0·510	0·023	0·493	0·058	0·486	0·063
1·0	JC	1·104	0·074	0·981	0·087	0·951	0·138
	K2P	1·017	0·076	1·025	0·118	0·968	0·146
	K3ST	1·020	0·075	1·027	0·118	0·975	0·149
	TK	1·022	0·075	1·033	0·124	0·976	0·150
	GIN	1·030	0·075	1·004	0·105	0·981	0·151
	TN	1·023	0·074	1·013	0·100	0·973	0·146
	G4H	1·021	0·073	1·033	0·133	0·983	0·156
2·0	JC	1·789	0·287	1·115	0·661	1·490	0·445
	K2P	1·828	0·320	1·072	0·674 (1)	1·557	0·512
	K3ST	1·842	0·328	1·084	0·688 (1)	1·588	0·593 (1)
	TK	2·049	0·718	1·092	0·691 (1)	1·596	0·596 (1)
	GIN	1·819	0·315 (2)	1·184	0·719	1·456	0·557 (3)
	TN	1·817	0·295	1·152	0·680	1·548	0·483
	G4H	1·836	0·375	1·180	0·722	1·592	0·530
0·125	JC	0·126	0·005	0·125	0·005	0·122	0·003
	K2P	0·126	0·005	0·126	0·004	0·124	0·004
	K3ST	0·126	0·005	0·127	0·004	0·124	0·004
	TK	0·127	0·005	0·118	0·005	0·121	0·005
	GIN	0·127	0·005	0·112	0·004	0·120	0·003
	TN	0·127	0·005	0·127	0·005	0·124	0·004
	G4H	0·127	0·005	0·127	0·004	0·124	0·004
0·25	JC	0·249	0·012	0·242	0·009	0·246	0·015
	K2P	0·249	0·013	0·245	0·009	0·247	0·015
	K3ST	0·250	0·013	0·249	0·009	0·247	0·015

True distance	Method	Substitution schemes: JC $\bar{\delta}$	JC σ_δ	K2P $\bar{\delta}$	K2P σ_δ	K3ST $\bar{\delta}$	K3ST σ_δ
0·25	TK	0·252	0·016	0·243	0·005	0·235	0·016
	GIN	0·251	0·013	0·250	0·010	0·243	0·014
	TN	0·251	0·013	0·252	0·009	0·250	0·016
	G4H	0·251	0·013	0·250	0·010	0·250	0·017
0·5	JC	0·496	0·018	0·469	0·025	0·512	0·028
	K2P	0·501	0·019	0·477	0·027	0·515	0·030
	K3ST	0·501	0·019	0·494	0·028	0·516	0·029
	TK	0·504	0·017	0·480	0·034	0·499	0·032
	GIN	0·503	0·018	0·497	0·027	0·520	0·031
	TN	0·505	0·019	0·502	0·029	0·532	0·031
	G4H	0·505	0·018	0·496	0·025	0·532	0·030
1·0	JC	0·909	0·126	0·724	0·170	0·898	0·143
	K2P	0·918	0·128	0·759	0·184	0·906	0·144
	K3ST	0·919	0·128	0·813	0·211	0·912	0·148
	TK	0·940	0·137	0·784	0·204	0·901	0·154
	GIN	0·939	0·136	0·838	0·231	0·961	0·162
	TN	0·935	0·131	0·819	0·210	0·975	0·168
	G4H	0·938	0·136	0·831	0·228	0·972	0·167
2·0	JC	1·302	0·460	1·183	0·407	1·311	0·388
	K2P	1·349	0·494	1·244	0·462 (1)	1·297	0·389 (1)
	K3ST	1·353	0·496	1·488	0·601	1·323	0·408 (1)
	TK	1·437	0·539	1·519	0·751 (1)	1·350	0·436 (1)
	GIN	1·469	0·566	1·324	0·607 (3)	1·494	0·583 (2)
	TN	1·378	0·500	1·491	0·669	1·681	0·745
	G4H	1·434	0·550	1·551	0·696	1·618	0·605

patterns while $\delta \leq 1$, whereas GIN method seems better for $\delta > 1$. Also, they noted that the frequency of cases in which TN is inapplicable is much lower than with GIN or TK methods. Because of that and the simplicity of TN formula, they considered TN method preferable overall. Our results agree in general with the above.

As can be seen from Table 4, G4H formula works fairly well when applied on these substitution schemes, it has showed to be applicable in all cases, and taking into account all figures, G4H estimates are the most regular in approximation to the true value of δ. As noted by Gojobori *et al.* (1982), the standard deviation increases with increasing δ, and not always the best estimates are obtained when applying a method on its appropriate substitution scheme.

Our results with schemes RAND0, RAND1 and RAND2, in which pseudo random matrices of substitution are used, show that G4H formula gives in general good estimates for $\delta \leq 1$, when $\delta = 2$ all methods tested give underestimates of the true δ. Inapplicable cases occur only in applying TK and GIN methods.

The frequency of inapplicable cases that have occurred in all our computer simulation is quite low in comparison with that obtained in other computer experi-

TABLE 5

Means ($\bar{\delta}$) and S.DS (σ_δ) of δ. Substitution schemes used are RAND0, RAND1 and RAND2. N = number of replications. The number of inapplicable cases is given in brackets

True distance	Method	Substitution schemes: RAND0 ($N=10$) $\bar{\delta}$	RAND0 ($N=10$) σ_δ	RAND1 ($N=50$) $\bar{\delta}$	RAND1 ($N=50$) σ_δ	RAND2 ($N=50$) $\bar{\delta}$	RAND2 ($N=50$) σ_δ
0·125	JC	0·127	0·007	0·122	0·008	0·121	0·010
	K2P	0·126	0·007	0·124	0·008	0·122	0·010
	K3ST	0·126	0·007	0·124	0·008	0·122	0·010
	TK	0·124	0·007	0·119	0·008	0·127	0·011
	GIN	0·124	0·008	0·120	0·007	0·127	0·012
	TN	0·128	0·007	0·124	0·008	0·124	0·010
	G4H	0·128	0·008	0·124	0·008	0·124	0·010
0·25	JC	0·255	0·013	0·247	0·011	0·236	0·021
	K2P	0·254	0·013	0·250	0·011	0·237	0·021
	K3ST	0·254	0·013	0·250	0·011	0·237	0·021
	TK	0·251	0·013	0·242	0·012	0·249	0·023
	GIN	0·252	0·012	0·245	0·011	0·255	0·025
	TN	0·258	0·013	0·253	0·012	0·247	0·023
	G4H	0·259	0·011	0·252	0·011	0·247	0·022
0·5	JC	0·512	0·023	0·475	0·058	0·440	0·058
	K2P	0·511	0·023	0·483	0·060	0·442	0·058
	K3ST	0·513	0·022	0·484	0·061	0·443	0·059
	TK	0·510	0·023	0·472	0·059	0·475	0·064
	GIN	0·512	0·018	0·480	0·061	0·494	0·074
	TN	0·526	0·025	0·494	0·063	0·477	0·067
	G4H	0·525	0·026	0·494	0·065	0·482	0·067
1·0	JC	0·939	0·086	0·852	0·130	0·734	0·153
	K2P	0·939	0·084	0·874	0·139	0·739	0·155
	K3ST	0·945	0·086	0·881	0·141	0·743	0·157
	TK	0·951	0·089	0·866	0·141	0·824	0·191
	GIN	0·962	0·099	0·902	0·155	0·887	0·242
	TN	0·981	0·095	0·907	0·148	0·834	0·193
	G4H	0·995	0·098	0·926	0·166	0·878	0·219
2·0	JC	1·626	0·353	1·295	0·361	1·015	0·318
	K2P	1·627	0·360	1·371	0·405	1·040	0·335
	K3ST	1·651	0·376	1·389	0·409	1·054	0·347
	TK	1·578	0·329 (2)	1·404	0·433 (1)	1·182	0·434 (3)
	GIN	1·804	0·465 (2)	1·460	0·474 (5)	1·123	0·440 (13)
	TN	1·835	0·515	1·447	0·437	1·243	0·464
	G4H	1·745	0·382	1·516	0·495	1·372	0·524

ments carried out by Gojobori *et al.* (1982) and Tajima & Nei (1984), see also Tajima (1985). In Gojobori *et al.*'s computer simulation many inapplicable cases were produced when their six parameter model of nucleotide substitution was used. In order to analyze the frequency of inapplicable cases we have conducted a computer simulation by using the substitution model of Gojobori *et al.* (1982) (see Table 3).

TABLE 6

Results of computer simulation [means ($\bar{\delta}$) and S.DS (σ_δ) of δ] using the substitution scheme of Gojobori et al. (1982) *are given in column A. Results reported by Gojobori et al.* (1982) *are given in column B, results reported by Tajima & Nei* (1984) *are given in column C.* L = *number of nucleotide pairs.* N = *number of replications. The number of inapplicable cases is given in brackets*

		A		B		C	
		$L=640$	$N=50$	$L=500$	$N=16$	$L=500$	$N=50$
True distance	Method	$\bar{\delta}$	σ_δ	$\bar{\delta}$		$\bar{\delta}$	σ_δ
1·0	JC	0·636	0·188	0·78		0·78	0·05
	K2P	0·653	0·196				
	K3ST	0·673	0·206				
	TK	0·775	0·284 (1)	1·07			
	GIN	0·758	0·258 (2)	1·06 (4)			
	TN	0·762	0·251			0·97	0·08
	G4H	0·779	0·265				
		$L=640$	$N=50$	$L=500$	$N=32$	$L=500$	$N=32$
		$\bar{\delta}$	σ_δ	$\bar{\delta}$		$\bar{\delta}$	σ_δ
2·0	JC	0·745	0·253	1·23		1·22	0·12
	K2P	0·771	0·272				
	K3ST	0·797	0·289				
	TK	0·899	0·359 (2)	1·53 (8)			
	GIN	0·909	0·407 (7)	2·20 (23)			
	TN	0·945	0·387			2·02	0·60 (3)
	G4H	0·991	0·428				

Our results are given in Table 6 which includes results reported by Gojobori *et al.* (1982) and Tajima & Nei (1984). We have found a much lower frequency of inapplicable cases through our computer simulation, we also note that estimates of δ by different methods give greater underestimates when tested on our simulation experiment. Both circumstances could be due to differences in the simulation procedure used.

How Useful are the Models?

An interesting pattern emerges from our simulation experiments: All models give similar estimates in all cases, and this common estimate improves as the number of changes decreases. All the models fail when $\delta > 1{\cdot}5$. It is also surprising that the effectiveness of the formulae is not dependent of the substitution scheme used, as it is clearly shown by the results obtained with the random schemes. Thus, it can be deduced that additional restrictions to the general reversible model do not improve the approximation of the estimates to the true δ, even if the appropriate substitution scheme is used. Therefore, our results suggest that (1) there exists an insurmountable limit to the precision of our estimates of the true number of substitutions occurring

in DNA evolution and (2) this limit is mainly dependent on the magnitude of the evolutionary distance, instead of the accuracy of the model used. Since the utility of the correction formulae decreases as evolutionary distance increases, our results allow to question the usefulness of any model based on the four hypotheses of the general model.

Conclusions

The values of the nucleotide substitution rates may be obtained only for models having at most six independent parameters. A remarkable case are the reversible models, for which eqn (7) holds.

However, the expected number of nucleotide substitutions per site may be obtained, by using eqn (4), for models having any number of parameters if the model has a given property described in the text. In particular, the expected number of substitutions may be obtained for commutative and for equifrequency models.

In any case, the utility of any correction formula based on the G4H model decreases as evolutionary distance increases.

We are most grateful to Dr M. Nei, who, several years ago, encouraged us with his comments to an early version of this paper. This work was financed in part by the DGICYT of the Spanish government (PB87-0881).

REFERENCES

BLAISDELL, B. E. (1985). A method for 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. *J. molec. Evol.* **22**, 69-81.

GOJOBORI, T., ISHII, K. & NEI, M. (1982). Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J. molec. Evol.* **18**, 414-423.

JUKES, T. H. & CANTOR, C. R. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism* (Munro, H. N., ed.) pp. 21-123. New York: Academic Press.

HARDISON, R. C., BUTLER III, E. T., LACY, E., MANIATIS, T., ROSENTHAL, N. & EFSTRATIADIS, A. (1979). The structure and transcription of four linked rabbit β-like globin genes. *Cell* **18**, 1279-1285.

KIMURA, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. molec. Evol.* **16**, 111-120.

KIMURA, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. natn. Acad. Sci. U.S.A.* **78**, 454-458.

KONKEL, D. A., MAIZEL JR, J. V. & LEDER, P. (1979). The evolution and sequence comparison of two recently diverged mouse chromosomal β-globin genes. *Cell* **18**, 865-873.

LANAVE, C., PREPARATA, G., SACCONE, C. & SERIO, G. (1984). A new method for calculating evolutionary substitution rates. *J. molec. Evol.* **20**, 86-93.

NEI, M. (1987). *Molecular Evolutionary Genetics.* New York: Columbia University Press.

OLIVER, J. L., MARÍN, A. & MEDINA, J. R. (1989). SDSE: a software package to simulate the evolution of a pair of DNA sequences. *Comput. appl. Biosci.* (CABIOS) **5**, 47-50.

TAJIMA, F. (1985). Estimation of evolutionary distance at the DNA level. In: *Population Genetics and Molecular Evolution* (Ohta, T. & Aoki, K., eds) pp. 281-292. Tokyo: Japan Scientific Societies Press/Springer-Verlag.

TAJIMA, F. & NEI, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Molec. 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 pseudogenes. *Genetics* **98**, 641-657.

APPENDIX I

We note that;

$$k = \sum_{i=1}^{i=4} q_i \left[\sum_{i \neq j} r_{ij} \right] = - \sum_{i=1}^{i=4} q_i r_{ii} = -\mathrm{tr}\,[X(0)R],$$

where tr denotes the trace of the matrix, that is the sum of the elements of the main diagonal.

For any square matrix:

$$(A')^n = (A^n)'$$

and

$$(A' + B' + C' + \ldots) = (A + B + C + \ldots)',$$

so that;

$$\exp(R't) = I + R't + \ldots + (R')^n t^n / n! + \ldots$$
$$= [I + (Rt)' + \ldots + (R^n t^n / n!)' + \ldots] = [\exp(Rt)]' = P'(t),$$

so, from eqn (1), we have

$$X(0)^{-1} X(t) = X(0)^{-1} P'(t) X(0) P(t),$$

where $X(0)^{-1}$ is the inverse of $X(0)$, that is a matrix such that

$$X(0)X(0)^{-1} = X(0)^{-1}X(0) = I.$$

From eqn (3), we have;

$$X(0)^{-1}X(t) = X(0)^{-1}[\exp(R't)]X(0)[\exp(Rt)],$$

and since for any matrix function $f(ABA^{-1}) = Af(B)A^{-1}$, we have:

$$X(0)^{-1}X(t) = \exp[X(0)^{-1}R'tX(0)] \exp(Rt),$$

so that;

$$\log[X(0)^{-1}X(t)] = X(0)^{-1}R'tX(0) + Rt + F[X(0)^{-1}R'tX(0), Rt], \qquad (4)$$

where F is a polynomial in its matrix arguments whose trace is zero.

From eqn (A.1) we obtain;

$$X(0) \log[X(0)^{-1}X(t)] = R'tX(0) + X(0)Rt + X(0)F,$$

i.e.;

$$R'tX(0) + X(0)Rt = X(0) \log[X(0)^{-1}X(t)] - X(0)F,$$

which implies that

$$\delta = -2t\,\mathrm{tr}\,[X(0)R] = -t\,\mathrm{tr}\,[R'X(0) + X(0)R]$$
$$= -\mathrm{tr}\,[(X(0)) \log(X(0)^{-1}X(t))] + \mathrm{tr}\,[X(0)F],$$

where the second term is known if $\mathrm{tr}\,[X(0)F]$ is zero.

APPENDIX II

As for any two matrices A and B that commute, it holds that;

$$\log[\exp(A)\exp(B)] = A + B,$$

we have;

$$\log(\exp[X(0)^{-1}R'tX(0)]\exp[(Rt)]) = X(0)^{-1}R'tX(0) + Rt,$$

which implies that [see eqn (A.1)] in this case

$$\log[X(0)^{-1}X(t)] = X(0)^{-1}RtX(0) + Rt$$

so that F must be zero.

APPENDIX III

Since for any two square matrices $(AB)' = B'A'$, and from eqn (3):

$$X(0)P(t) = X(0)I + X(0)R + \ldots + X(0)R^n t^n/n! + \ldots,$$

$$P'(t)X(0) = IX(0) + R'X(0) + \ldots + (R')^n X(0)t^n/n! + \ldots,$$

and, for every n it holds that;

$$X(0)R^n = (R')^n X(0) = (R^n)'X(0)$$

because it holds for $n = 1$, and if it holds for $n = h - 1$

$$X(0)R^{h-1} = (R^{h-1})'X(0),$$

it also holds for $n = h$:

$$X(0)R^h = X(0)R^{h-1}R = (R^{h-1})'X(0)R = (R^{h-1})'R'X'(0)$$
$$= (RR^{h-1})'X(0) = (R^h)'X(0).$$

Then,

$$X(0)I + X(0)R + \ldots + X(0)R^n t^n/n! + \ldots$$
$$= IX(0) + R'X(0) + \ldots + (R')^n X(0)t^n/n! + \ldots$$

because every member of the first expression is equal to the correspondent member of the second one.