Copyright © 1995 by the Genetics Society of America

Estimating Substitution Rates in Ribosomal RNA Genes

Andrey Rzhetsky

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received May 8, 1995
Accepted for publication June 28, 1995

ABSTRACT

A model is introduced describing nucleotide substitution in ribosomal RNA (rRNA) genes. In this model, substitution in the stem and loop regions of rRNA is modeled with 16- and four-state continuous time Markov chains, respectively. The mean substitution rates at nucleotide sites are assumed to follow gamma distributions that are different for the two types of regions. The simplest formulation of the model allows for explicit expressions for transition probabilities of the Markov processes to be found. These expressions were used to analyze several 16S-like rRNA genes from higher eukaryotes with the maximum likelihood method. Although the observed proportion of invariable sites was only slightly higher in the stem regions, the estimated average substitution rates in the stem regions were almost two times as high as in the loop regions. Therefore, the degree of site heterogeneity of substitution rates in the stem regions seems to be higher than in the loop regions of animal 16S-like rRNAs due to presence of a few rapidly evolving sites. The model appears to be helpful in understanding the regularities of nucleotide substitution in rRNAs and probably minimizing errors in recovering phylogeny for distantly related taxa from these genes.

RIBOSOMAL RNA (rRNA) genes are used extensively for inferring evolutionary relationships among species because they allow for meaningful comparison of sequences from very distant taxa (e.g., see WOESE 1987; SIMON *et al.* 1993; WAINRIGHT *et al.* 1993; RAGAN *et al.* 1994; SMOTHERS *et al.* 1994) and are found in all nonviral organisms. Almost all of the previous phylogenetic analyses of rRNAs have been performed with either the maximum parsimony method, which does not incorporate any explicit model of nucleotide substitution, or with various model-based tree-making methods under JUKES and CANTOR's (1969) or KIMURA's (1980) models of nucleotide substitution. These models are built on a number of restrictive assumptions that are often violated in real data. Even though the results of the published phylogenetic analyses are in most cases supported by independent morphological, paleontological or biogeographical data, the robustness of the tree-making methods is not generally guaranteed if the underlying mathematical model does not fit the data. While the computational cost of application of parameter-rich models to real data may become prohibitively high, it is desirable to match specific properties of data with the model to minimize the risk associated with using an oversimplified model.

There are a considerable number of existing mathematical models that are applicable to rRNA genes to varying degrees. Most of these models assume independence of nucleotide sites and homogeneity of substitution rates across sequences but allow for various patterns of nucleotide substitution (for review see NEI 1987; RODRIGUEZ *et al.* 1990). Some of them allow for rate variation across sites (*e.g.*, see GOLDING 1983; NEI and GOJOBORI 1986; JIN and NEI 1990; LI *et al.* 1990; TAKAHATA 1991; YANG 1993), and a few recently proposed models allow for nonindependence of nucleotide substitution in certain sites within the same gene (SHÖNIGER and VON HAESELER 1994; TILLIER 1994; MUSE 1995). However, none of these models was designed specifically to describe nucleotide substitution in rRNA genes, where both site dependence and variability of substitution rates across the sequences are known to be important (VAWTER and BROWN 1993; KUMAR and RZHETSKY 1995). Below I describe a mathematical model designed with the objective of incorporating the most important features of rRNA evolution while keeping the number of parameters small. Although this paper is written from the perspective of estimating evolutionary distances between extant sequences, it also provides some new analytical expressions that will be useful in phylogenetic analysis of rRNA sequences with the maximum likelihood method.

To explain the choice of assumptions and parameters used in the model, I shall first briefly review the set of known functional constraints that affect the pattern of nucleotide substitution in rRNA genes.

FUNCTIONAL CONSTRAINTS ON NUCLEOTIDE SUBSTITUTION IN RRNA GENES

The most prominent characteristic of all rRNAs is their highly conserved secondary structure defined by comple-

Address for correspondence: Andrey Rzhetsky, 328 Mueller Laboratory, Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, PA 16802-5301. E-mail: aur1@psuvm.psu.edu

mentary regions within the same molecule. Two types of intramolecule interactions are known in rRNAs: short-range pairings creating hairpins and long-range pairings that order several short-range pairings into more sophisticated structures (JAMES *et al.* 1988; GUTELL 1992). About half of nucleotide sites [*e.g.*, 57% in the case of 16S-like rRNAs of vertebrates (VAWTER and BROWN 1993) and 48% in the case of 23S rRNA of *Escherichia coli* (NOLLER *et al.* 1981)] do not participate in the Watson-Crick interactions with other sites in the same RNA strand. Hence, a mature rRNA molecule comprises both single-stranded (loop) and double-stranded (stem) regions that presumably differ in their modes of substitution.

A significant proportion of sites that are critically important for a normal function of a ribosome reside in the loop regions of rRNAs. For example, in 16S rRNA of *E. coli*, the single-stranded regions are responsible for association of ribosome subunits, binding of peptidyl tRNA, recognition of Shine-Dalgarno sequence in bacterial mRNAs, and interaction with the protein factor IF3 (*e.g.*, see WOESE *et al.* 1980; GLOTZ *et al.* 1981). Since the vast majority of mutations in the highly constrained sites are deleterious, such sites display little or no variation (GUTELL *et al.* 1985). Other sites in the loop regions are more variable, and their functional constraints, although not well understood, seem to be less demanding. The pattern of nucleotide substitution appears to vary among different loop regions and, at least in the case of 16S-like rRNAs of vertebrates (VAWTER and BROWN 1993), there seems to be no consistent transition/transversion substitution bias.

The primary function of the stem regions is thought to be the maintenance of the secondary structure of the molecule. Therefore, mutations occurring in a stem region are individually deleterious if they destabilize a functionally important structure, but fitness can be restored, when a compensatory mutation occurs that reestablishes the pairing potential (JAMES *et al.* 1988). Among all non-Watson-Crick nucleotide pairs, the U · G (uracil · guanine) pair appears to be the least deleterious (JAMES *et al.* 1988) and in some cases even at selective advantage (ROUSSET *et al.* 1991). This pair is not rare in the stem regions of functional rRNAs. For example, in human 18S rRNA 15% of site pairs in stem regions are occupied by U · G.

The rates of nucleotide substitution may not be constant across sites in the stem regions. One reason to expect the rate heterogeneity is the so-called distance effect that was described by STEPHAN and KIRBY (1993) for the stem regions of *Adh* precursor messenger RNAs in Drosophila but was not directly studied in rRNA genes. The regularity found by STEPHAN and KIRBY (1993) implied that substitution rates in the stem region tend to decrease as the physical distance between the paired sites becomes larger. To explain this observation STEPHAN and KIRBY (1993) referred to KIMURA's (1985) model describing fixation of compensatory mutations in a haploid population. Although this model is not directly applicable to nucleotide substitution in rRNA stem regions, it indicates qualitatively that the per site substitution rate in the stem regions should decrease as recombination distance between two paired sites becomes larger provided that nucleotide substitution in the double-stranded regions occurs through alternation of neutral (Watson-Crick pairs) and deleterious (non-canonical pairs) states. Another factor that is likely to contribute to variation in the substitution rates in the stem regions is a set of constraints imposed by interaction of rRNAs with ribosomal proteins (*e.g.*, see WOESE *et al.* 1980; NOLLER *et al.* 1981; AAGAARD and DOUTHWAITE, 1994).

An approximate analysis of the distribution of substitution rates across rRNA sequences can be performed with the maximum parsimony method (FITCH 1971; UZZEL and CORBIN 1971; WAKELEY 1993). If substitution rates do not vary among sites and nucleotide substitution along branches of the true tree is governed by a Markov chain in continuous time, the actual numbers of nucleotide substitutions per site must follow a Poisson distribution. This assumption is roughly tested as follows (UZZEL and CORBIN 1971). First, a preliminary tree topology is obtained from the data of interest assuming that substitution rates are homogeneous across sites. Next, basing on this preliminary tree, the number of changes required at each site is computed using parsimonious reconstruction of ancestral states at each interior node. Theoretical curves are then fitted to the observed distribution of the numbers of nucleotide substitutions per site. A frequency distribution obtained in the described way from both stem and loop regions of eukaryotic 16S-like rRNAs (KUMAR and RZHETSKY 1995) was clearly different from a Poisson distribution but was sufficiently close to a negative binomial distribution with the same mean and variance. The negative binomial distribution of substitution counts is expected under the assumption that nucleotide substitution in each site is governed by a Poisson process, where the rate is itself a random variable that follows a gamma distribution (see UZZEL and CORBIN 1971; GOLDING 1983; JIN and NEI 1990; TAKAHATA 1991; YANG 1993). For the present study it is important to note that the observed distribution cannot be satisfactorily fitted by a weighted sum of two Poisson distributions (which should be expected if the stem and the loop regions evolved at different rates but substitution rates within each class of sites were homogeneous), since the composite distribution with the same mean and variance (7.5 and 80, respectively) must be bimodal. In contrast, the shape of the observed distribution is consistent with a weighted sum of two negative binomial distributions. This indicates that a model assuming that substitution rates in the stem and the loop regions follow two different gamma distributions is not incompatible with 16S-like rRNA data in an obvious way.

Now we are in position to introduce parameters and assumptions defining the model.

MODEL

Assumptions and parameters: Consider a set of m present-day rRNA sequences of total length l. Our ob-

jective is to estimate parameters of the evolutionary process that generated these sequences under the following assumptions.

1. The genes under analysis are homologous and related by an (unknown) true tree.
2. The homologous sites in different genes are known and the only source of change in the sequence evolution is nucleotide substitution.
3. The sites of the present-day genes can be unambiguously classified into two groups, n pairs of sites in the stem regions and $(l - 2n)$ sites in the loop regions.
4. The patterns of nucleotide substitution are different for the stem and the loop regions. Each pair of interacting sites in a double-stranded region and each site in a single-stranded region are modeled by the first-order continuous-time Markov processes with 16 and four discrete states, respectively.
5. The Markov process in each pair of interacting sites in the stem regions is independent of the Markov processes in other sites. Furthermore, each site in the single-stranded regions evolves independently of other sites, and the 16- and four-state Markov processes are not correlated.
6. The Markov process describing substitution in the ith pair of sites in a stem region along the jth branch of the true tree is defined as follows. To describe the heterogeneity of substitution rates across sites, we follow a number of authors (*e.g.*, see GOLDING 1983; NEI and GOJOBORI 1986; JIN and NEI 1990; LI *et al.* 1990; TAKAHATA 1991; YANG 1993) and assume that the rate of the Markov process in the ith site pair of the sequence alignment is determined by a random variable X_i that is sampled from a gamma distribution with density function $f(x; a_S)$. Subscripts S and L will indicate hereafter the quantities related to the stem and loop regions, respectively. Notation $f(z; a)$ stands for

$$f(z; a) = \begin{cases} \dfrac{a^a z^{a-1} e^{-az}}{\Gamma(a)}, & z \geq 0, \\ 0, & z < 0. \end{cases} \quad (1)$$

Density function 1 is obtained by replacing both shape and scale parameters of the conventional gamma density function with a single parameter, a, where $0 < a \leq \infty$. This one-parameter gamma distribution has mean 1 and variance $1/a$. We assume that random variables X_i's for different pairs of sites ($i = 1, \ldots, n$) are independently and identically distributed and therefore the parameter of the gamma distribution, a_S, can be estimated from data. The probability of having pair of nucleotides (or dinucleotide for brevity) v replaced with dinucleotide w ($v \neq w$) within a small time interval $[\tau, \tau + \Delta\tau]$ in the ith pair of sites is defined by

$$[\mathbf{Q}_S]_{vw} X_i \rho_j(\tau) \Delta\tau + [\text{small terms including } (\Delta\tau)^2, (\Delta\tau)^3, \text{ etc.}]. \quad (2)$$

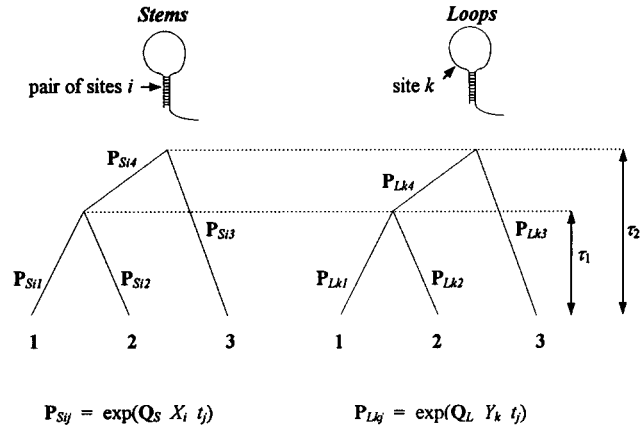


FIGURE 1.—Hypothetical true tree for three extant rRNA sequences and matrices of transition probabilities along each branch shown separately for the stem and the loop regions.

Here $\rho_j(\tau)$ is a nonnegative function specific to the jth branch, and $\mathbf{Q}_S$ is 16×16 rate matrix independent of τ. All off-diagonal elements in matrix $\mathbf{Q}_S$ are nonnegative and the diagonal elements are chosen to ensure that elements in each row sum to zero. Matrix $\mathbf{Q}_S$ summarizes information about the pattern of dinucleotide substitution, whereas function $\rho_j(\tau)$ specifies fluctuations in the mean substitution rate throughout evolutionary history. In a special case, when $\rho_j(\tau)$ is a constant independent of j, the Markov process defined by (2) is homogeneous in time. In more general case we can consider a new time parameter t_j defined as the integral of $\rho_j(\tau)$ over time interval between endpoints of the jth branch. Then the matrix of transition probabilities is expressed by $\mathbf{P}_{Sij} = \exp(\mathbf{Q}_S X_i t_j)$, where subscripts i and j refer to the ith pair of sites and the jth branch of the true tree, respectively (see Figure 1).

7. The four-state Markov process is defined by analogy with the 16-state process. In this case 16×16 matrix $\mathbf{Q}_S$ in (2) is replaced with 4×4 matrix $\mathbf{Q}_L$, and the site-specific rate in the kth site of a loop region is assumed to be determined by random variable Y_k (see Figure 1). As before, we assume that all Y_k's are independent and follow the same gamma distribution with density $f(y; a_L)$ as defined by (1). The intensity functions, $\rho_j(\tau)$'s, are postulated to be the same for both types of Markov processes for each branch of the true tree.
8. Markov processes of the two types allow for existence of equilibrium distributions defined by (1×16) vector $\boldsymbol{\pi}_S$ and (1×4) vector $\boldsymbol{\pi}_L$. Vector $\boldsymbol{\pi}_S$ satisfies

$$\boldsymbol{\pi}_S \mathbf{Q}_S = \mathbf{0}, \quad \sum_{h=1}^{16} \pi_{Sh} = 1, \quad (3)$$

where $\mathbf{0}$ is a zero vector, and vector $\boldsymbol{\pi}_L$ is defined by analogy with (3).
9. Distributions of states in the most recent common ancestor of m extant sequences are assumed identical to the equilibrium distribution vectors $\boldsymbol{\pi}_S$ and $\boldsymbol{\pi}_L$ for the stem and the loop regions, respectively.

Note that it is possible to estimate products, $(\mathbf{Q}_S\, t_j)$ and $(\mathbf{Q}_L\, t_j)$, but not $\mathbf{Q}_S$, $\mathbf{Q}_L$, and t_j separately. Nevertheless, it is convenient to keep the "redundant" parameters in the following derivation since this allows the Markov processes that are not homogeneous in time to be treated as time homogeneous. The length of the jth branch (δ_{Sj} and δ_{Lj}, for the stem and the loop regions, respectively) is defined by the mean number of nucleotide substitutions per site along this branch, where

$$\delta_{Sj} = -\tfrac{1}{2} \sum_{i=1}^{16} \boldsymbol{\pi}_{Si}[\mathbf{Q}_S]_{ii} t_j, \quad \text{and}$$

$$\delta_{Lj} = -\sum_{k=1}^{4} \boldsymbol{\pi}_{Lk}[\mathbf{Q}_L]_{kk} t_j. \quad (4)$$

The evolutionary distance between a pair of extant rRNA genes is then defined by the sum of branch lengths along the shortest path connecting these sequences in the true tree. For example, in the case of present-day sequences 1 and 2 in Figure 1, the distance between their stem regions (d_S) is given by

$$d_S = -\tfrac{1}{2} \sum_{i=1}^{16} \boldsymbol{\pi}_{Si}[\mathbf{Q}_S]_{ii}(t_1 + t_2). \quad (5)$$

Finally, we can introduce the total distance, d_T, between two present-day sequences by $d_T = d_S + d_L$, where d_L is the distance between the loop regions. In the following we shall consider certain simplifying assumptions about matrices $\mathbf{Q}_S$ and $\mathbf{Q}_L$.

The loop regions: In the most general case matrix $\mathbf{Q}_L$ may comprise 12 different parameters. Since the pattern of nucleotide substitution in the single-stranded regions seems to vary across the sequence (see VAWTER and BROWN 1993), for the moment we shall use the simplest possible matrix $\mathbf{Q}_L$ defined by

$$[\mathbf{Q}_L]_{ij} = \begin{cases} -3\epsilon, & i = j, \\ \epsilon, & i \neq j, \end{cases} \quad (6)$$

which specifies the model described first by JUKES and CANTOR (1969). The length of the jth branch of the true tree [see (4)] under this model is defined by

$$\delta_{Lj} = 3\epsilon t_j. \quad (7)$$

Omitting the derivation of transitional probabilities under this model, consider the estimation of the distance (d_L) between the loop regions of two present-day sequences, say sequences 1 and 2 in Figure 1. It is impossible to estimate a pair of parameters, the evolutionary distance (d_L) and the gamma parameter (a_L) simultaneously from a pair of present-day sequences that provide only one independent observation (the proportion of differences) under the Jukes-Cantor model. However, assuming that the value of a_L is known, the maximum likelihood estimate of d_L can be obtained by solving the following equation with respect to $\hat{d}_L$ (see GOLDING 1983; NEI and GOJOBORI 1986; JIN and NEI 1990; LI *et al.* 1991).

$$\hat{p}_L = \tfrac{3}{4}\{1 - [a_L/(a_L + \tfrac{4}{3}\hat{d}_L)]^{a_L}\}, \quad (8)$$

where $\hat{p}_L$ is the observed proportion of differences between the loop regions of sequences 1 and 2.

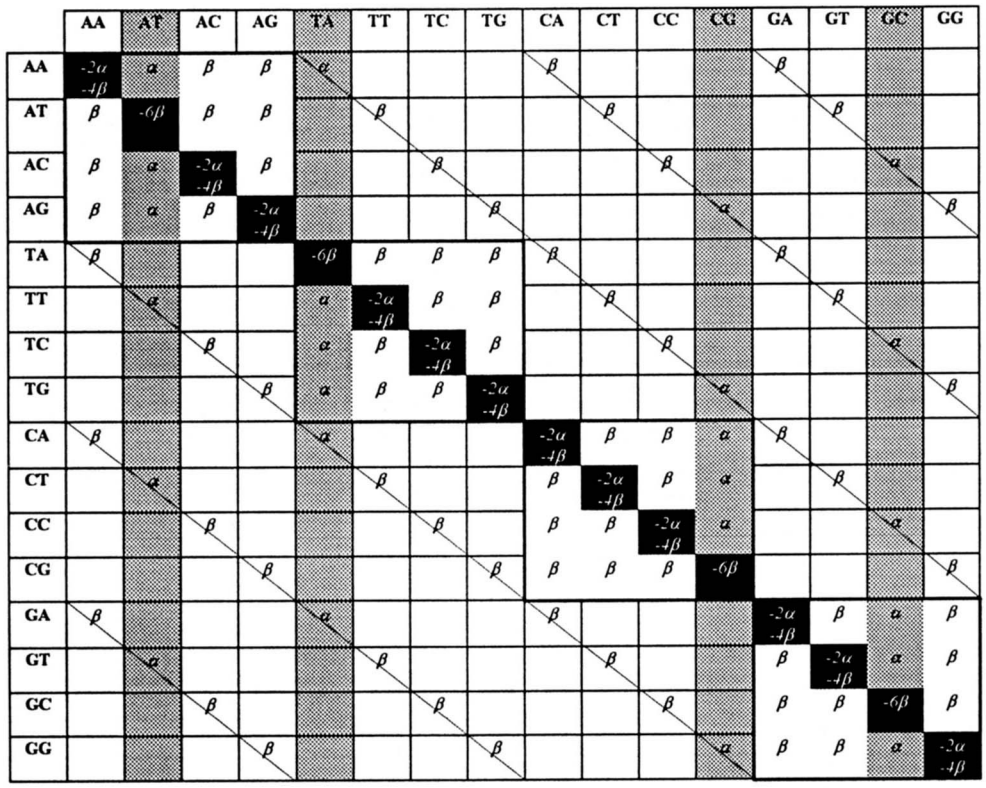
The stem regions: Matrix $\mathbf{Q}_S$ combines information about patterns of both mutation and selection in the stem regions of rRNAs. By virtue of the requirement that entries in each row of this matrix sum to zero, in the most general case matrix $\mathbf{Q}_S$ includes $16 \times 16 - 16 = 240$ independent parameters. In reality it is hardly practical to use a model with hundreds of parameters, especially when a moderate amount of data is available. Therefore, some assumptions about entries of $\mathbf{Q}_S$ must be introduced. First, assuming that mutations occur independently at different sites and that their number for a fixed time interval follows a Poisson distribution, one can set all instantaneous rates of two-substitution transitions (such as A-T → C-G) to zero. [This assumption was used by SHÖNIGER and VON HAESELER (1994) and MUSE (1995), but not by TILLIER (1994).] This brings the number of parameters in $\mathbf{Q}_S$ down to 98, which is still rather large. One can further decrease the number of parameters to 16 by setting all nonzero entries within each column of the matrix to the same value (see SHÖNIGER and VON HAESELER 1994). This can be considered to describe a case when mutation changes any nucleotide to any other with equal probability but selective values of the 16 nucleotide pairs are different. [MUSE (1995) incorporated somewhat different symmetry of the substitution process into his model. In contrast, TILLIER (1994) reduced the overall number of parameters by postulating that only six dinucleotide states out of 16 possible are allowed in the stem regions.] Finally, assuming that dinucleotides can be classified into three equivalence classes (the Watson-Crick pairs, $T \cdot G$ and $G \cdot T$ pairs, and the rest of noncanonical pairs) and that dinucleotides within each group are selectively identical, we can decrease the total number of parameters to three. Therefore, matrix $\mathbf{Q}_S$ considered in this paper is a special case of the SHÖNIGER-VON HAESELER matrix. The off-diagonal elements of this matrix are defined as follows.

$$[\mathbf{Q}_S]_{ij} = \begin{cases} \alpha, & j \text{ is a Watson-Crick pair, one} \\ & \text{difference between } i \text{ and } j, \\ \gamma, & j \text{ is } GT \text{ or } TG \text{ pair,} \\ & \text{one difference between } i \text{ and } j, \\ \beta, & j \text{ is any other noncanonical pair,} \\ & \text{one difference between } i \text{ and } j, \\ 0, & \text{two differences between } i \text{ and } j. \end{cases} \quad (9)$$

The diagonal elements are chosen to ensure that row sums are zero. Biologically reasonable values of α, β, and γ should probably satisfy the inequality

$$\alpha \geq \gamma \geq \beta > 0. \quad (10)$$

Matrix $\mathbf{Q}_S$



$$\alpha \geq \beta \geq 0$$

FIGURE 2.—Two-parameter matrix $\mathbf{Q}_S$. Empty cells correspond to zero entries of the matrix. Borders and shadows are used to emphasize symmetric properties of the matrix.

Before considering the distance estimation, let us highlight several important properties of the substitution model defined by matrix 9. First, in the case $\alpha = \beta = \gamma$ the 16-state Markov model reduces to JUKES and CANTOR's (1969) model with four independent nucleotide states. Second, the rate of nucleotide substitution is higher in the case $\alpha = \beta = \gamma$ than in the case $\alpha > \gamma > \beta$. The third useful observation is that matrix $\mathbf{Q}_S$ in (9) defines a time-reversible Markov process, *i.e.*, matrix $\text{diag}\{\boldsymbol{\pi}_S\}\mathbf{Q}_S$ is real symmetrical (see KEILSON 1979), where $\text{diag}\{\boldsymbol{\pi}_S\}$ is a square matrix with elements of vector $\boldsymbol{\pi}_S$ on the main diagonal and all off-diagonal elements equal to zero.

Estimating evolutionary distance between the stem regions: Here we shall consider only a special case of matrix 9 where $\beta = \gamma$ (see Figures 2 and 3) and there are just two groups of identical states, paired (A-T, T-A, C-G, and G-C) and unpaired (all other dinucleotides), because the derivation of the corresponding equations under the three-parameter model turned out to be very cumbersome.

To obtain an expression for estimating d_S from the observed differences between present-day sequences, one has to go through a number of well-defined computational steps. The first step is to determine the matrix of transition probabilities between 16 states, $\mathbf{P}_S(t, x)$, where $[\mathbf{P}_S(t, x)]_{vw}$ is the probability of having dinucleotide v replaced by dinucleotide w after amount of time t given that the relative substitution rate at the pair of sites considered is equal to x. The symmetry of the substitution process (see Figure 3) allows this matrix to be determined in a simple and elegant way (see APPENDIX A). Namely, we have

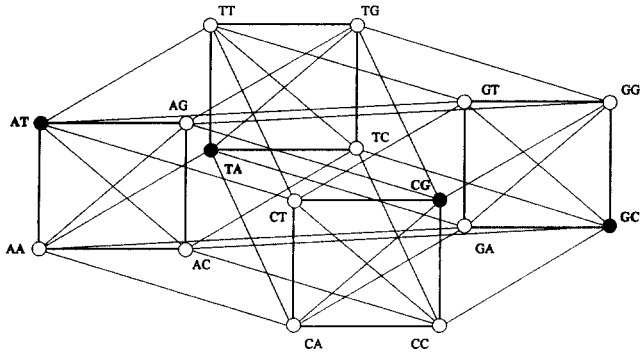
$$\mathbf{P}_S(t, x) = \mathbf{R}_1 + \exp\{-2(\alpha + \beta)tx\}\mathbf{R}_2 + \exp\{-4\beta tx\}\mathbf{R}_3 + \exp\{-2(\alpha + 3\beta)tx\}\mathbf{R}_4, \quad (11)$$

where matrices, $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$ are as shown in Figure 4, and the equilibrium frequencies of paired and unpaired states are given by

$$\pi_p = \frac{\alpha}{4(\alpha + 3\beta)}, \quad \text{and} \quad \pi_u = \frac{\beta}{4(\alpha + 3\beta)}, \quad (12)$$

respectively.

Once $\mathbf{P}_S(t, x)$ is determined, it is easy to compute probabilities of observing dinucleotide v and dinucleotide w in homologous sites of two present-day sequences (say, sequences 1 and 2 in Figure 1) given that the relative substitution rate at this site is equal to x. We denote the matrix of these probabilities by $\mathbf{X}_S(t_1, t_2, x)$, where



1. Paired state —No differences→ Paired state (AT → AT)
2. Paired state —Two differences→ Paired state (AT → CG)
3. Paired state —One difference→ Unpaired state (AT → AA)
4. Paired state —Two differences→ Unpaired state (AT → CA)
5. Unpaired state —No differences→ Unpaired state (AA → AA)
6. Unpaired state —One difference→ Unpaired state (AG → AA)
7. Unpaired state —Two differences→ Unpaired state (AG → CA)
8. Unpaired state —One difference→ Paired state (AG → AT)
9. Unpaired state —Two differences→ Paired state (AG → GC)

FIGURE 3.—Graph of adjacency of dinucleotide states in matrix $\mathbf{Q}_S$ in Figure 2. There are two types of identical states: paired (●) and unpaired (○). Only dinucleotides one difference apart are adjacent. The symmetry of this graph suggests that the entries of matrix $\mathbf{P}_S$ in (14) can be classified into nine equivalence groups.

$$\mathbf{X}_S(t_1, t_2, x) = [\mathbf{P}_S(t_1, x)]^t \operatorname{diag}\{\boldsymbol{\pi}_S\}\mathbf{P}_S(t_2, x), \quad (13)$$

(*e.g.*, see TAVARÉ 1986). It is then possible to find the weighted average of matrix 14 over all possible relative rates, x's, using density function 1.

$$\mathbf{X}_S = \int_0^\infty f(x; a_S)\mathbf{X}_S(t_1, t_2, x)\,dx. \quad (14)$$

The resulting matrix $\mathbf{X}_S$ is shown in Figure 5. This matrix can be used to obtain expressions for estimating d_S from the observed frequencies of dinucleotide pairs between two present-day sequences. The most statistically efficient way to estimate parameters d_S and a_S from two extant sequences is to find the values of parameters that maximize the appropriate likelihood function. Noting that matrix $\mathbf{X}_S$ comprises 256 entries of only nine different types (*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, and *I*) and omitting a constant multiplier, we can express the logarithm of the likelihood function as follows:

$$\begin{aligned}\log L = {} & 12\hat{A}\log(A) + 48\hat{B}\log(B) + 48\hat{C}\log(C) \\ & + 12\hat{D}\log(D) + 48\hat{E}\log(E) \\ & + 24\hat{F}\log(F) + 48\hat{G}\log(G) \\ & + 4\hat{H}\log(H) + 12\hat{I}\log(I), \quad (15)\end{aligned}$$

where *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, and *I* are parameters defined in Figure 5, and the same letters with hats denote the observed frequencies of dinucleotide pairs corresponding to each probability value. For example, $12\hat{A}$ stands for sum of the observed frequencies of all pairs of identical noncanonical dinucleotides (*e.g.*, A·A/A·A) in the present-day sequences, see Figure 5. Unfortunately, the values of d_S and a_S that maximize function 15 could not be recovered analytically and one has to rely on a numerical technique.

It is still possible to derive various (nonmaximum likelihood) closed-form expressions that provide consistent estimate ($\hat{d}_S$) of d_S from real data assuming that the value of a_S is known. We shall consider here only one expression of this kind introducing parameter $S = 48B + 48G + 4H + 12I$ (see Figure 5). One can verify that

$$\begin{aligned}d_S = 24(\pi_p + \pi_u)\beta(t_1 + t_2) = {} & -\frac{12a_S(\pi_p + \pi_u)\pi_u}{(\pi_p + 3\pi_u)} \\ & \times \left\{1 - \left[2 + \frac{\pi_p}{3\pi_u} - \frac{S}{48\pi_p\pi_u}\right]^{-1/a_S}\right\}, \quad (16)\end{aligned}$$

where d_S is expressed in terms of parameters π_p, π_u, S, and a_S. Parameters π_p, π_u, and S can be directly estimated from two present-day sequences, and parameter a_S is assumed to be known. Hence, $\hat{d}_S$ can be computed by substituting parameters in the right side of (16) with their estimates, and the variance of this estimate can be evaluated by the "delta-technique" (see KENDALL and STUART 1958).

Considering the properties of expression 16, observe that in the case $(t_1 + t_2) = \infty$ we have $S = 16\pi_p(\pi_p + 6\pi_u)$, and $d_S = \infty$. Therefore, (16) is inapplicable for estimating d_S when $\hat{S}$ is greater than $16\hat{\pi}_p(\hat{\pi}_p + 6\hat{\pi}_u)$. Equation 16 can be used to derive corresponding formula for d_S under the assumption that substitution rates are homogeneous across stem regions, *i.e.*, parameter a_S tends to infinity. In the following sections we shall consider estimation of parameters a_S and a_L from real sequences and compare various ways of computing distances between the stem regions.

Estimating parameters of the model: To get a feeling about the values of parameters that can be expected in real data analysis, I analyzed a set of eukaryotic 16S-like rRNA genes. This set included sequences representing mammals (*Homo sapiens*), arthropods (*Acyrthosiphon pisum*), nematodes (*Caenorabditis elegans* and *Strongyloides stercoralis*), and fungi (*Saccharomyces cerevisiae*). Because the evolutionary relationships of these taxa seem to be noncontroversial, one can concentrate on estimation of the model parameters for a predetermined tree instead of discriminating among alternative phylogenies first. The aligned sequences were retrieved from the Ribosomal Database Project (LARSEN *et al.* 1993) (see legend to Figure 6). To identify the boundaries of the stem and loop regions, previously published secondary structures (GUTELL *et al.* 1985; GUTELL 1991) were used.

Matrix $\mathbf{R}_1$

π_u	π_p	π_u	π_u	π_p	π_u	π_u	π_u	π_u	π_u	π_u	π_p	π_u	π_u	π_p	π_u
π_u	π_p	π_u	π_u	π_p	π_u	π_u	π_u	π_u	π_u	π_u	π_p	π_u	π_u	π_p	π_u
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
π_u	π_p	π_u	π_u	π_p	π_u	π_u	π_u	π_u	π_u	π_u	π_p	π_u	π_u	π_p	π_u

Matrix $\mathbf{R}_2$

1/4		1/8	1/8		-1/4	-1/8	-1/8	1/8	-1/8			1/8	-1/8		
1/8		1/4	1/8		-1/8	1/8			-1/8	1/8		-1/8	-1/4		-1/8
1/8		1/8	1/4		-1/8		1/8	-1/8	-1/4	-1/8			-1/8		1/8
-1/4		-1/8	-1/8		1/4	1/8	1/8	-1/8	1/8			-1/8	1/8		
-1/8		1/8			1/8	1/4	1/8	-1/8		1/8		-1/8	-1/8		-1/8
-1/8			1/8		1/8	1/8	1/4	-1/4	-1/8	-1/8		-1/8			1/8
1/8			-1/8		-1/8	-1/8	-1/4	1/4	1/8	1/8		1/8			-1/8
-1/8		-1/8	-1/4		1/8		-1/8	1/8	1/4	1/8			1/8		-1/8
		1/8	-1/8			1/8	-1/8	1/8	1/8	1/4		-1/8	-1/8		-1/4
1/8		-1/8			-1/8	-1/4	-1/8	1/8		-1/8		1/4	1/8		1/8
-1/8		-1/4	-1/8		1/8	-1/8			1/8	-1/8		1/8	1/4		1/8
		-1/8	1/8			-1/8	1/8	-1/8	-1/8	-1/4		1/8	1/8		1/4

Matrix $\mathbf{R}_3$

a	*b*			*b*	*a*					-*a*	-*b*			-*b*	-*a*
a	*c*	*a*	*a*	-*b*	*a*	-*a*	-*a*	-*a*	*a*	-*a*	-*b*	-*a*	*a*	-*b*	-*a*
	b	*a*		-*b*			-*a*	-*a*			-*b*		*a*	*b*	
	b		*a*	-*b*		-*a*			*a*		*b*	-*a*		-*b*	
a	-*b*	-*a*	-*a*	*c*	*a*	*a*	*a*	*a*	-*a*	-*a*	-*b*	*a*	-*a*	-*b*	-*a*
a	*b*			*b*	*a*					-*a*	-*b*			-*b*	-*a*
	-*b*		-*a*	*b*		*a*			-*a*		-*b*	*a*		*b*	
	-*b*	-*a*		*b*			*a*	*a*			*b*		-*a*	-*b*	
	-*b*	-*a*		*b*			*a*	*a*			*b*		-*a*	-*b*	
	b		*a*	-*b*		-*a*			*a*		*b*	-*a*		-*b*	
-*a*	-*b*			-*b*	-*a*					*a*	*b*			*b*	*a*
-*a*	-*b*	-*a*	*a*	-*b*	-*a*	-*a*	*a*	*a*	*a*	*a*	*c*	-*a*	-*a*	-*b*	*a*
	-*b*		-*a*	*b*		*a*			-*a*		-*b*	*a*		*b*	
	b	*a*		-*b*			-*a*	-*a*			-*b*		*a*	*b*	
-*a*	-*b*	*a*	-*a*	-*b*	-*a*	*a*	-*a*	-*a*	-*a*	*a*	-*b*	*a*	*a*	*c*	*a*
-*a*	-*b*			-*b*	-*a*					*a*	*b*			*b*	*a*

$$a = \frac{\beta}{4(\alpha+\beta)}, \qquad b = \frac{\alpha}{4(\alpha+\beta)}, \qquad c = \frac{3\alpha}{4(\alpha+\beta)}$$

FIGURE 4.—Spectral matrices $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$, each of size 16 × 16. All empty cells correspond to zero entries of the matrices. Borders and shadows are used to emphasize the symmetric properties of matrices.

Two data sets for the stem region were used. The large data set (387 pairs of sites from four animal sequences) included all sites that were likely to belong to the stem regions, and in the small data set (343 pairs of sites from the same sequences) all sites where alignment was equivocal were deleted. The data set for the loop regions included 804 sites from the five sequences. All sites containing insertions, deletions or ambiguous characters were excluded from these data sets.

The maximum likelihood method used in these analyses was implemented as described by FELSENSTEIN (1981), who introduced an algorithm of evaluating a likelihood function for unrooted trees under a time-reversible model of nucleotide substitution, and by YANG (1993), who considered new aspects of this computation associated with adjusting this model to variation of substitution rates across sites. In the case of the 16-state model it was convenient to express matrices $\mathbf{R}_3$, and $\mathbf{R}_4$ in Figure 4 in terms of parameters π_p and π_u by introducing a new parameter, $R = a/\beta = \pi_p/\pi_u$. In this case the entries of matrices $\mathbf{R}_3$ and $\mathbf{R}_4$ were easier to compute because parameter R can be directly estimated from the extant sequences. The estimates of parameters a_S (from the large data set) and a_L obtained in the

Matrix R_4

d	e	f	f	e	g	h	h	f	h	i	j	f	h	j	i
k	l	k	k	j	k	i	i	i	k	i	j	i	k	j	i
f	e	d	f	j	h	f	i	i	h	f	j	h	g	e	h
f	e	f	d	j	h	i	f	h	g	h	e	i	h	j	f
k	j	i	i	l	k	k	k	k	i	i	j	k	i	j	i
g	e	h	h	e	d	f	f	h	f	i	j	h	f	j	i
h	j	f	i	e	f	d	f	h	i	f	j	g	h	e	h
h	j	i	f	e	f	f	d	g	h	h	e	h	i	j	f
f	j	i	h	e	h	h	g	d	f	f	e	f	i	j	h
h	e	h	g	j	f	i	h	f	d	f	e	i	f	j	h
i	j	f	h	j	i	f	h	f	f	d	e	h	h	e	g
i	j	i	k	j	i	i	k	k	k	k	l	i	i	j	k
f	j	h	i	e	h	g	h	f	i	h	j	d	f	e	f
h	e	g	h	j	f	h	i	i	f	h	j	f	d	e	f
i	j	k	i	j	i	k	i	i	k	i	j	k	k	l	k
i	j	h	f	j	i	h	f	h	h	g	e	f	f	e	d

$$d = \frac{3\alpha^2 + 10\alpha\beta + 5\beta^2}{4(\alpha^2 + 4\alpha\beta + 3\beta^2)}, \qquad e = -\frac{\alpha(\alpha + 2\beta)}{2(\alpha^2 + 4\alpha\beta + 3\beta^2)},$$

$$f = -\frac{\alpha + 5\beta}{8(\alpha + 3\beta)}, \qquad g = \frac{\alpha^2 + 2\alpha\beta - \beta^2}{4(\alpha^2 + 4\alpha\beta + 3\beta^2)},$$

$$h = \frac{\alpha + \beta}{8(\alpha + 3\beta)}, \qquad i = \frac{\beta^2}{2(\alpha^2 + 4\alpha\beta + 3\beta^2)},$$

$$j = \frac{\alpha\beta}{2(\alpha^2 + 4\alpha\beta + 3\beta^2)}, \qquad k = -\frac{\beta(\alpha + 2\beta)}{2(\alpha^2 + 4\alpha\beta + 3\beta^2)},$$

$$l = \frac{3\beta(\alpha + 2\beta)}{2(\alpha^2 + 4\alpha\beta + 3\beta^2)}.$$

FIGURE 4.—*Continued*

analysis as well as estimated branch lengths of the tree are shown in Figure 6. All five sequences were used in analysis for the loop regions, but only four animal sequences were used in estimation of parameters for the stem region. (This is because optimization of the likelihood function for the four sequences under the "continuous gamma" model required about 36 hr with a SUN SPARC 5 workstation. An analogous computation for five sequences would require ~256 times as much as for the four sequences.) Results obtained for the loop regions (Figure 6B) were independently verified with a computer program BASEMLG (YANG 1993) provided by Z. YANG.

The outcome of these analyses turned out to be interesting in two respects. First, the estimate of a_S was approximately twice as small as the estimate of a_L (see Figure 6). The estimate of a_S from the small data set was a little more extreme (0.268) than from the large data set (0.287). Second, the estimated branch lengths for the stem region tree turned out to be on average twice as large as corresponding estimates for the loop region tree (see Figure 6). Although the branch lengths for the stem-region tree were somewhat shorter when estimated from the small data set (data not shown), they were still longer than the branch lengths for the loop region tree (see also VAWTER and BROWN 1993). Curiously, the proportion of invariable sites in the four animal genes was somewhat larger in the stem regions (0.62 and 0.69 for the large and the small data sets, respectively) than in the loop regions (0.59). This suggests that the stem regions include a few sites that evolve very rapidly but the rest of the sites in the stem regions change rather slowly. In other words, the selective pressure in the loop regions seems to be higher but distributed more evenly than in the stem regions.

A crude analysis of distribution of the fast-evolving sites along the double-stranded regions suggested that the "distance effect" (STEPHAN and KIRBY 1993) was not dominating in the evolution of 16S-like rRNA genes. More precisely, there were no clear-cut correlation between the relative substitution rate in a pair of sites and the physical distance between these sites (data not shown). Instead, the fast-changing sites tended to be situated in the middle of long stems, in the optional helices that are absent in some species (see WOESE *et al.* 1983) and at the boundaries of short conservative stems, *i.e.*, in the regions where the impact of nucleotide substitution on the rRNA secondary structure seems to be the least harmful. To support these observations statistically an additional data analysis using a larger data set needs to be performed. The reader is also addressed to the paper by G. B. GOLDING (1994) where distribution of selective strength across stem regions of 16S-like rRNA genes was estimated with the maximum likelihood method under a population-genetics model.

Application of (16) to estimating distances between the real sequences indicated that the variances of the estimates were large. To verify this I compared alternative estimators of evolutionary distances between the stem regions in computer simulation assuming that the value of a_S is fixed (0.287) and known. Results of this comparison were somewhat surprising (see Table 1). It was anticipated that the distance estimates obtained with (16) would be less efficient than those derived by maximizing the likelihood function 15. The actual difference in the efficiency was striking: although both estimators are consistent (see the last row of Table 1), the variances of estimates obtained with (16) in some cases were about two orders of magnitude higher than the variances of the maximum-likelihood estimates. In addition, the estimates obtained from relatively short simulated sequences displayed positive bias that was small for the maximum-likelihood estimates but quite large for the estimates computed with (16). Therefore, the practical importance of (16) for analysis of short sequences is rather limited, say, to finding the starting values for maximizing the likelihood function in 15. The maximum likelihood estimates of distances obtained from a set of sequences were rather close to those obtained from a pair of sequences (see Table 2), although the pairwise estimates appeared to be consistently larger.

I also experimented with estimating parameters a_S

Matrix X_S

	AA	AT	AC	AG	TA	TT	TC	TG	CA	CT	CC	CG	GA	GT	GC	GG
AA	*A*	*B*	*C*	*C*	*B*	*D*	*E*	*E*	*C*	*E*	*F*	*G*	*C*	*E*	*G*	*F*
AT	*B*	*H*	*B*	*B*	*I*	*B*	*G*	*G*	*G*	*B*	*G*	*I*	*G*	*B*	*I*	*G*
AC	*C*	*B*	*A*	*C*	*G*	*E*	*C*	*F*	*F*	*E*	*C*	*G*	*E*	*D*	*B*	*E*
AG	*C*	*B*	*C*	*A*	*G*	*E*	*F*	*C*	*E*	*D*	*E*	*B*	*F*	*E*	*G*	*C*
TA	*B*	*I*	*G*	*G*	*H*	*B*	*B*	*B*	*B*	*G*	*G*	*I*	*B*	*G*	*I*	*G*
TT	*D*	*B*	*E*	*E*	*B*	*A*	*C*	*C*	*E*	*C*	*F*	*G*	*E*	*C*	*G*	*F*
TC	*E*	*G*	*C*	*F*	*B*	*C*	*A*	*C*	*E*	*F*	*C*	*G*	*D*	*E*	*B*	*E*
TG	*E*	*G*	*F*	*C*	*B*	*C*	*C*	*A*	*D*	*E*	*E*	*B*	*E*	*F*	*G*	*C*
CA	*C*	*G*	*F*	*E*	*B*	*E*	*E*	*D*	*A*	*C*	*C*	*B*	*C*	*F*	*G*	*E*
CT	*E*	*B*	*E*	*D*	*G*	*C*	*F*	*E*	*C*	*A*	*C*	*B*	*F*	*C*	*G*	*E*
CC	*F*	*G*	*C*	*E*	*G*	*F*	*C*	*E*	*C*	*C*	*A*	*B*	*E*	*E*	*B*	*D*
CG	*G*	*I*	*G*	*B*	*I*	*G*	*G*	*B*	*B*	*B*	*B*	*H*	*G*	*G*	*I*	*B*
GA	*C*	*G*	*E*	*F*	*B*	*E*	*D*	*E*	*C*	*F*	*E*	*G*	*A*	*C*	*B*	*C*
GT	*E*	*B*	*D*	*E*	*G*	*C*	*E*	*F*	*F*	*C*	*E*	*G*	*C*	*A*	*B*	*C*
GC	*G*	*I*	*B*	*G*	*I*	*G*	*B*	*G*	*G*	*G*	*B*	*I*	*B*	*B*	*H*	*B*
GG	*F*	*G*	*E*	*C*	*G*	*F*	*E*	*C*	*E*	*E*	*D*	*B*	*C*	*C*	*B*	*A*

$$A = \pi_u(\pi_u, 1/4, a, d)\,\mathbf{e}, \quad B = \pi_u(\pi_p, 0, b, e)\,\mathbf{e} = \pi_p(\pi_u, 0, a, k)\,\mathbf{e},$$

$$C = \pi_u(\pi_u, 1/8, 0, f)\,\mathbf{e}, \quad D = \pi_u(\pi_u, -1/4, a, g)\,\mathbf{e}, \quad E = \pi_u(\pi_u, -1/8, 0, h)\,\mathbf{e},$$

$$F = \pi_u(\pi_u, 0, -a, i)\,\mathbf{e}, \quad G = \pi_u(\pi_p, 0, -b, j)\,\mathbf{e} = \pi_p(\pi_u, 0, -a, i)\,\mathbf{e},$$

$$H = \pi_p(\pi_p, 0, c, l)\,\mathbf{e}, \quad I = \pi_p(\pi_p, 0, -b, j)\,\mathbf{e},$$

$$\mathbf{e}^t = \left(1, \left[\frac{a_S}{a_S + 2(\alpha+\beta)(t_1+t_2)}\right]^{a_S}, \left[\frac{a_S}{a_S + 4\beta(t_1+t_2)}\right]^{a_S}, \left[\frac{a_S}{a_S + 2(\alpha+3\beta)(t_1+t_2)}\right]^{a_S}\right).$$

$$S = 4(12B + 12G + 3I + H).$$

FIGURE 5.—Matrix of the expected frequencies of dinucleotide pairs in the stem regions of sequences 1 and 2 in Figure 1. The values of parameters *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *j*, *k*, and *l* are as shown in Figure 4, and superscript *t* indicates transpose of a vector.

and d_S simultaneously from a pair of sequences. In contrast to the case described by YANG (1994), it was not impossible to obtain consistent estimates of the parameters from two extant sequences under the 16-state model. However, computer simulation indicated that these estimates had a tendency to have both large sampling variances and considerable positive biases when simulated sequences were not very long (data not shown). Therefore, estimation of parameter a_S from a pair of sequences does not seem to be practical with real data.

DISCUSSION

The model considered in this paper is built on a number of assumptions that can be tested in subsequent analyses. One important assumption is that fluctuations in substitution rates in the stem and the loop regions are synchronized, *i.e.*, the ratio between the expected lengths of the analogous branches in the stem and the loop region trees (δ_{Si}/δ_{Li}) remains constant over all branches of the true tree. If this assumption is violated, one cannot meaningfully combine pairwise distances computed for the stem and the loop regions into a total distance, d_T. Since, here, the values of gamma-parameters (a_S and a_L) were assumed constant for all lineages, it is worthwhile to check whether estimated values of these parameters differ significantly among taxonomic groups. Another set of biologically important hypotheses is associated with sets of symmetrical restrictions on the entries of rate matrices ($\mathbf{Q}_S$ and $\mathbf{Q}_L$) and possible variation of these restrictions among evolutionary lineages. All these hypotheses can be tested in the framework of the maximum likelihood analysis.

The practical use of the above model in real data analysis can be substantially complicated by uncertainties in alignment of nucleotide sequences. When applying this model to analysis of real genes one has to exactly specify the position of paired sites in the double-stranded regions and the boundaries between the stem and the loop regions throughout the set of sequences in addition to usual identification of homologous sites in different genes. All nucleotide sites that cannot be unambiguously aligned and assigned to either class should be excluded from further analysis. The number of excluded sites can be particularly large when distantly related rRNA genes representing, say, different eukaryotic kingdoms are compared, because both size and position of the stem and the loop regions tend to vary across taxa (GUTELL 1992) and are not well known for certain species. Hopefully, efforts to overcome these complications will be rewarded by insights into regularities of nucleotide substitution in rRNA genes. In addition, application of the appropriate mathematical models in phylogenetic analysis of rRNA genes may help to

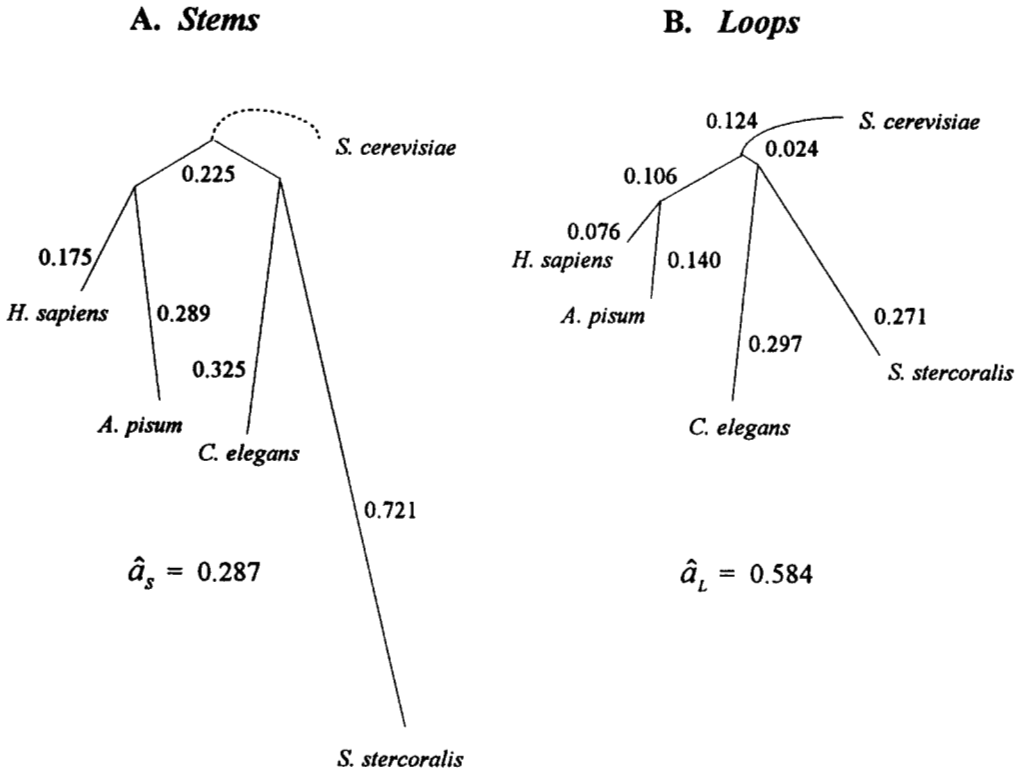


FIGURE 6.—Branch lengths and parameters of gamma distributions estimated from the stem regions (A) and the loop regions (B) of several 16S-like rRNA sequences with the maximum-likelihood method. In both cases the likelihood functions were maximized for the same predetermined tree topology that appears to be almost certainly true from a taxonomist's point of view. The taxa represented in the tree include mammals (*H. sapiens*), arthropods (*A. pisum*), nematods (*C. elegans* and *S. stercoralis*), and fungi (*S. cerevisiae*). The branch lengths of the trees are shown in terms of the average number of nucleotide substitutions per site. The model parameters were estimated from 387 site pairs for the stem regions (four sequences) and 804 sites from the loop regions (five sequences). The small subunit rRNA genes used in the analysis were derived from the Ribosomal Database Project on the anonymous ftp server at the University of Illinois in Urbana, Illinois updated on August 7, 1993 (LARSEN *et al.* 1993).

reduce risk of making systematic errors in estimating the actual relationships among taxa (see TILLIER and COLLINS 1995; VON HAESELER and SCHÖNIGER 1995).

An interesting and somewhat counterintuitive property of the model for the stem regions considered in this paper (see matrix $\mathbf{Q}_S$ in Figure 2) is that the numbers of nucleotide substitutions occurring in two interacting sites in a stem follow two independent distributions [this is also true for the model by S. V. MUSE (1995), see APPENDIX B]. One might suspect that by analogy with (8) for the loop regions, the distance between the stem regions can be expressed just in terms of the expected proportion of differences between the stem regions, p_S. This turns out to be indeed the case. Combining the expected frequencies of dinucleotide pairs having a difference at site 1 (and any configuration in

TABLE 1

Comparison of the distance estimates calculated with (16) with those obtained by maximizing the likelihood function 15

	d_S = 0.05		d_S = 0.25	
n	$\hat{d}_S$ (16)	$\hat{d}_S$ (15)	$\hat{d}_S$ (16)	$\hat{d}_S$ (15)
250	0.0794 (0.0112)	0.0505 (0.0002)	0.3929 (0.2860)	0.2533 (0.0027)
500	0.0622 (0.0031)	0.0502 (0.0001)	0.3049 (0.0483)	0.2515 (0.0013)
1000	0.0559 (0.0012)	0.0501 (0.0000)	0.2759 (0.0161)	0.2506 (0.0005)
1500	0.0537 (0.0007)	0.0501 (0.0000)	0.2655 (0.0088)	0.2505 (0.0003)
∞	0.05	0.05	0.25	0.25

The average values of estimates and their variances (shown in parentheses) were computed in 1000 simulation replications for each value of the expected distance (d_S) and the number of paired sites under analysis (n). The values of $\hat{d}_S$ for $n = \infty$ were obtained by setting the observed frequencies of dinucleotide pairs to their expected values. In all computations, the true values of parameters a_S (=0.287) and R (=α/β = 8.19) used in generating the extant sequences were assumed to be known.

TABLE 2

Comparison of distance estimates computed with the maximum likelihood method from pairs of sequences with those computed from the set of four sequences

Sequences[a]	$\hat{d}_S$ (15)	$\hat{d}_S$ (four sequences)
1 and 2	0.526	0.464
1 and 3	0.885	0.725
1 and 4	1.277	1.121
2 and 3	0.992	0.839
2 and 4	1.442	1.235
3 and 4	1.135	1.046

[a] The sequences are enumerated as follows: 1, *H. sapiens;* 2, *A. pisum;* 3, *C. elegans;* 4, *S. stercoralis.*

site 2), we find that $p_S = 24(B + C + D/2 + 2E + F + 2G + I/2)$, see Figure 5, and

$$p_S = \frac{3}{4} - 6\pi_u\left[\frac{a_S}{a_S + (R+1)d_S/[12(\pi_u + \pi_p)]}\right]^{a_S} - 3(\pi_u + \pi_p)\left[\frac{a_S}{a_S + d_S/[6(\pi_u + \pi_p)]}\right]^{a_S}. \quad (17)$$

Letting $\alpha = \beta$ (in which case $\pi_u = \pi_p = 1/16$ and $R = 1$) and substituting parameters p_S and d_S with their estimates, one can transform (17) into (8). Although estimates of d_S obtained with (17) (assuming that a_S is known) are not very efficient, this expression has certain appeal in not requiring knowledge about the exact position of paired sites within the stem regions. It should be emphasized that the independence of two distributions does not imply that paired sites in the stem regions evolve completely independently and, hence, that the rate of nucleotide substitution in the stem regions can be equivalently estimated under a four-state Jukes-Cantor-like model. In contradistinction to the usual four-state Markov process, the waiting times between two subsequent substitutions at the same site in a stem region under the current model follow two different exponential distributions rather than one, depending on the type of nucleotide pair occupying the paired sites.

The model for the stem regions considered in this paper appears to be analytically tractable and simple in implementation due to relatively small number of parameters. Nevertheless, computation of the likelihood function for a few sequences requires considerable time. Although it is possible to invoke techniques to significantly reduce the computational complexity (YANG 1994), a point where a model becomes computationally unfeasible is usually reached quickly as the number of sequences under analysis increases. Therefore, to reconstruct phylogenetic trees from sizable data under multiparameter models of nucleotide substitution it might be reasonable to combine distance-matrix tree-making algorithms with efficient estimation of evolutionary distances between sequences. The efficiency of the distance estimation can be increased by a careful selection of the distance estimator or using more than two sequences at a time for estimating pairwise distances between them. Being statistically consistent and relatively fast, the distance matrix methods may provide a reasonable compromise between computational complexity and statistical efficiency in phylogenetic inference from large data sets.

COMPUTER PROGRAM

An *ad hoc* computer program that was used in the data analysis described in this paper is available on request. The author could spend an additional amount of time and efforts to write a more general program if this is requested.

I am grateful to SPENCER V. MUSE for discussion and making his paper available before publication. Special thanks are due to M. NEI, T. SITNIKOVA, Z. YANG, S. B. HEDGES, S. KUMAR, M. STEEL, A. ZHARKIKH, S. V. MUSE, A. VON HAESELER, N. TAKAHATA and two anonymous referees for their comments and helpful suggestions on the earlier versions of the manuscript. ZIHENG YANG provided valuable advice and a computer subroutine for numerical minimization of a real-valued function in multidimensional space. ANDREY ZHARKIKH sacrificed a considerable amount of his time to reproduce equations presented in this paper. This study was supported by grants from National Institutes of Health and National Science Foundation to MASATOSHI NEI.

LITERATURE CITED

AAGAARD, C., and S. DOUTHWAITE, 1994 Requirement for a conserved, tertiary interaction in the core of 23S ribosomal RNA. Proc. Natl. Acad. Sci. USA **91:** 2989–2993.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

FITCH, W., 1971 Towards defining the course of evolution: minimum change for specific tree topology. Syst. Zool. **20:** 406–416.

GLOTZ, C., and R. BRIMACOMBE, 1980 An experimentally-derived model for the secondary structure of the 16S ribosomal RNA from *Escherichia coli.* Nucleic Acids Res. **8:** 2377–2395.

GOLDING, G. B., 1983 Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. **1:** 125–142.

GOLDING, G. B., 1994 Using maximum likelihood to infer selection from phylogenies, pp. 126–139 in *Non-neutral Evolution. Theories and Molecular Data,* edited by G. B. GOLDING. Chapman and Hall, New York.

GUTELL, R. R., 1992 Evolutionary characteristics of 16S and 23S rRNA structures, pp. 243–309 in *The Origin and Evolution of the Cell,* edited by H. HARTMAN and K. MATSUNO. World Scientific, Singapore.

GUTELL, R. R., B. WEISER, C. R. WOESE and H. F. NOLLER, 1985 Comparative anatomy of 16S-like ribosomal RNA. Prog. Nucleic Acids Res. Mol. Biol. **32:** 155–216.

JAMES, B. D., G. J. OLSEN, J. LIU and N. R. PACE, 1988 The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. Cell **52:** 19–26.

JIN, L., and M. NEI, 1990 Limitation of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7:** 82–102.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism,* Vol. 3, edited by H. N. MUNRO. Academic Press, New York.

HUNTER, J. J., 1983 *Mathematical Techniques of Applied Probability. Vol. 1. Discrete Models: Basic Theory.* Academic Press, New York.

KEILSON, J., 1979 *Markov Chain Models—Rarity and Exponentiality.* Springer-Verlag, New York.

KENDALL, G., and A. STUART, 1958 *The Advanced Theory of Statistics,* Vol. 1. Hafner, New York.

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., 1985 The role of compensatory neutral mutations in molecular evolution. J. Genet. **64:** 7–19.

KUMAR, S., and A. RZHETSKY, 1995 Evolutionary relationships of eukaryotic kingdoms. J. Mol. Evol. (in press).

LARSEN, N., G. J. OLSEN, B. L. MAIDAK, M. J. MCCAUGHEY, R. OVERBREEK *et al.*, 1993 The ribosomal database project. Nucleic Acids Res. **21** (Supplement): 3021–3023.

LI, W.-H., M. GOUY, P. M. SHARP, C. O'HUIGIN and Y.-W. YANG, 1990 Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla and Carnivora and molecular clocks. Proc. Natl. Acad. Sci. USA **87:** 6703–6707.

MUSE, S. V., 1995 Evolutionary analysis of DNA sequences subject to constraints on secondary structure. Genetics **139:** 1429–1439.

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

NEI, M., and T. GOJOBORI, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

NOLLER, H. F., J. KOP, V. WHEATON, J. BROSIUS, R. R. GUTELL *et al.*, 1981 Secondary structure model for 23S ribosomal RNA. Nucleic Acids Res. **9:** 6167–6189.

RAGAN, M., C. J. BIRD, E. L. RICE, R. R. GUTELL, C. A. MURPHY *et al.*, 1994 A molecular phylogeny of the marine red algae (Rhodophyta) based on the nuclear small-subunit rRNA gene. Proc. Natl. Acad. Sci. USA **91:** 7276–7280.

RAO, C. R., 1973 *Linear Statistical Inference and Its Applications.* Wiley, New York.

RODRIGUEZ, F., J. L. OLIVER, A. MARTIN and J. R. MEDINA, 1990 The general stochastic model of nucleotide substitution. J. Theor. Biol. **142:** 485–501.

ROUSSET, F., M. PÉLANDAKIS and M. SOLIGNAC, 1991 Evolution of compensatory substitutions trough G·U intermediate state in Drosophila rRNA. Proc. Natl. Acad. Sci. USA **88:** 10032–10036.

SCHÖNIGER, M., and A. VON HAESELER, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phyl. Evol. **3:** 240–247.

SIMON, L., J. BOUSQUET, R. C. LÉVESQUE and M. LALONDE, 1993 Origin and diversification of endomycorrhizal fungi and coincidence with vascular land plants. Nature **363:** 67–69.

SMOTHERS, J. F., C. D. VON DOHLEN, L. H. SMITH Jr. and R. D. SPALL, 1994 Molecular evidence that the myxozoans protists are metazoans. Science **265:** 1719–1721.

STEPHAN, W., and D. A. KIRBY, 1993 RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics. **135:** 97–103.

TAKAHATA, N., 1991 Overdispersed molecular clock at the major histocompatibility complex loci. Proc. R. Soc. Lond. B **243:** 13–18.

TAVARÉ, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences. **17:** 57–86.

TILLIER, E. R. M., 1994 Maximum likelihood with multiparameter models of substitution. J. Mol. Evol. **39:** 409–417.

TILLIER, E. R. M., and R. A. COLLINS, 1995 Neighbor-joining and maximum likelihood with RNA sequences: addressing interdependence of sites. Mol. Biol. Evol. **12:** 7–15.

UZZELL, T., and K. W. CORBIN, 1971 Fitting discrete probability distribution to evolutionary events. Science **253:** 1503–1507.

VAWTER, L., and W. M. BROWN, 1993 Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics **134:** 597–608.

VON HAESELER, A., and M. SCHÖNIGER, 1995 Ribosomal RNA phylogeny derived from a correlation model of sequence evolution, pp. 330–338 in *From Data to Knowledge,* edited by W. GAUL and D. PFEFFER. Springer, Berlin.

WAINRIGHT, P. O., G. HINKLE, M. L. SOGIN and S. K. STICKEL, 1993 Monophyletic origins of the metazoa: an evolutionary link with fungi. Science **260:** 340–342.

WAKELEY, J., 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

WOESE, C. R., 1987 Bacterial evolution. Microbiol. Rev. **51:** 221–271.

WOESE, C. R., L. J. MAGRUM, R. GUPTA, R. B. SIEGEL, D. A. STAHL *et al.*, 1980 Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. Nucleic Acids Res. **8:** 2275–2293.

WOESE, C. R., R. GUTELL, R. GUPTA and H. F. NOLLER, 1983 Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. Microbiol. Rev. **47:** 621–669.

YANG, Z., 1993 Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:** 1396–1401.

YANG, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

ZWEIB, C., and R. BRIMACOMBE, 1980 Localization of a series of intra-RNA cross-links by ultraviolet irradiation of *Escherichia coli* 30S ribosomal subunits. Nucleic Acids Res. **8:** 2397–2411.

Communicating editor: N. TAKAHATA

APPENDIX A: DERIVATION OF TRANSITION PROBABILITIES FOR THE 16-STATE MARKOV PROCESS

The explicit form of $\mathbf{P}_S(t, x)$ can be found from spectral decomposition of matrix $(\mathbf{Q}_S tx)$.

$$\mathbf{P}_S(t, x) = \exp\{\mathbf{Q}_S tx\} = \sum_{k=1}^{16} \exp(\lambda_k tx)\mathbf{u}_k\mathbf{v}_k, \quad \text{(A1)}$$

where λ_k, $\mathbf{u}_k$, and $\mathbf{v}_k$, are the kth eigenvalue of matrix $\mathbf{Q}_S$ and associated left (row) and right (column) eigenvectors, respectively (*e.g.*, see HUNTER 1983). To compute $\exp\{\mathbf{Q}_S tx\}$ with (A1) we start with finding the eigenvalues of matrix $\mathbf{Q}_S$ in Figure 2. Only four out of 16 eigenvalues of $\mathbf{Q}_S$ turn out to be distinct, namely

$$\lambda_1 = 0, \quad \lambda_2 = \lambda_3 = \lambda_4 = -2(\alpha + \beta),$$
$$\lambda_5 = \lambda_6 = \lambda_7 = -4\beta,$$
$$\text{and } \lambda_8 = \lambda_9 = \cdots = \lambda_{16} = -2(\alpha + 3\beta). \quad \text{(A2)}$$

It is clear from (3) that the equilibrium distribution vector $\boldsymbol{\pi}_S$ satisfies the equation for the left eigenvector $(\mathbf{u}_1)$ associated with eigenvalue $\lambda_1 = 0$. Therefore, we can choose

$$\mathbf{u}_1 = (\pi_u, \pi_p, \pi_u, \pi_u, \pi_p, \pi_u, \pi_u, \pi_u, \pi_u, \pi_u, \pi_p, \pi_u, \pi_u, \pi_p, \pi_u), \quad \text{(A3)}$$

where π_u and π_p are given in (12). Corresponding right eigenvector, $\mathbf{v}_1$, satisfying equation $\mathbf{u}_1\mathbf{v}_1 = 1$, is given by

$$\mathbf{v}_1^t = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1). \quad \text{(A4)}$$

The straightforward calculation of the remaining 15 pairs of left and right eigenvectors of matrix $\mathbf{Q}_S$ would require a repulsive amount of algebra. Fortunately, one can avoid this computation with the following trick. Noting that only four out of 16 eigenvalues of $\mathbf{Q}_S$ are distinct, we can rewrite (A1) in the form (11), where matrices, $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$ are defined by

$$\mathbf{R}_1 = \mathbf{u}_1\mathbf{v}_1, \quad \mathbf{R}_2 = \sum_{k=2}^{4} \mathbf{u}_k\mathbf{v}_k, \quad \mathbf{R}_3 = \sum_{k=5}^{7} \mathbf{u}_k\mathbf{v}_k,$$
$$\mathbf{R}_4 = \sum_{k=8}^{16} \mathbf{u}_k\mathbf{v}_k. \quad \text{(A5)}$$

Since matrix $\mathbf{R}_1$ is already known (see A3 and A4), it turns out to be much easier to determine matrices $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$ directly than compute the eigenvectors first. It will be shown below that these matrices can be readily found by using information about relationships of their entries.

The first set of relationships follows from the symmetrical properties of the model. Figure 3 indicates that there are at most nine different expressions for 256 conditional probabilities constituting entries of matrix $\mathbf{P}_S(t, x)$ in (A1). Therefore, each of the matrices $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$ has at most nine different entries, and we have to solve a system of 27 equations (rather than 256 $\times$ 3 equations as would be required in the absence of the symmetry). Figure 3 is helpful to identify positions of identical entries within these matrices.

Another set of constraints emerges from theory of spectral decomposition of a square matrix (*e.g.*, see HUNTER 1983) and the properties of orthogonal projectors (RAO 1973). Indeed, since our 16-state Markov process is irreducible (any state can be accessed from any other state), ergodic [there is a nonnegative vector $\boldsymbol{\pi}_S$ satisfying (3)] and time-reversible, matrices $\mathbf{u}_k\mathbf{v}_k$ in (A1) are all real-valued, idempotent, orthogonal, and each of them has rank one (see KEILSON 1979). It is then easy to obtain the following restrictions on matrices $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$.

$$\mathbf{R}_1[0]^k + \mathbf{R}_2[-2(\alpha + \beta)]^k + \mathbf{R}_3[-4\beta]^k + \mathrm{R}_4[-2(\alpha + 3\beta)]^k = \mathbf{Q}_S^k \quad \text{(A6)}$$

where $(k = 0, 1, 2, \ldots)$.

$$\mathbf{R}_i\mathbf{R}_j = \delta_{ij}\mathbf{R}_i \quad (i, j = 1, 2, 3, 4). \quad \text{(A7)}$$

$$\mathrm{rank}[\mathbf{R}_1] + \mathrm{rank}[\mathbf{R}_2] + \mathrm{rank}[\mathbf{R}_3] + \mathrm{rank}[\mathbf{R}_4] = \mathrm{rank}[\mathbf{I}] = 16, \quad \text{(A8)}$$

$$\mathrm{trace}[\mathbf{R}_i] = \mathrm{rank}[\mathbf{R}_i] = \mu_i, \quad \text{(A9)}$$

where $\mathbf{I}$ is 16 $\times$ 16 identity matrix, δ_{ij} is Kronecker's delta, μ_i is the algebraic multiplicity of the eigenvalue associated with the matrix $\mathbf{R}_i$, and trace[ ] and rank[ ] denote sum of diagonal elements and rank of the matrix in brackets, respectively. Combining (A2), (A8), and (A9), we obtain

$$\mathrm{trace}[\mathbf{R}_1] = 1, \quad \mathrm{trace}[\mathbf{R}_2] = \mathrm{trace}[\mathbf{R}_3] = 3, \quad \text{and } \mathrm{trace}[\mathbf{R}_4] = 9. \quad \text{(A10)}$$

Finally, the identity

$$\sum_{j=1}^{16} [\mathbf{R}_k]_{ij} = \begin{cases} 1, & k = 1, \\ 0, & k = 2, 3, 4 \end{cases} \quad \text{(A11)}$$

follows from (A1) and the observation that every row of matrix $\mathbf{P}_S(t, x)$ must sum to 1 for any $t \geq 0$.

Note that (A6) by itself allows for unambiguous identification of matrices $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$, and all the additional equations serve just to simplify computation. The explicit expressions for matrices $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$ are given in Figure 4 and can be readily verified to satisfy the above equations. The closed form expressions for conditional transition probabilities are immediately available due to (A1).

We can save quite a few additional algebraic operations by using the property of time-reversibility of the Markov processes under consideration. Indeed, instead of evaluating (13) directly, we can rewrite it as

$$\mathbf{X}_S(t_1, t_2, x) = \mathrm{diag}\{\boldsymbol{\pi}_S\}\mathbf{P}_S(t_1 + t_2, x) \quad \text{(A12)}$$

(TAVARÉ 1986). That is, we reverse direction of the process leading from the common ancestor of sequences 1 and 2 (see Figure 1) to sequence 1 and evaluate transition probabilities for amount of time $(t_1 + t_2)$ given that distribution of states at the starting point 1 is specified by vector $\boldsymbol{\pi}_S$. Next, we can get rid of variable x in expression for $\mathbf{X}_S(t_1, t_2, x)$ [see (14)] noting that matrices $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$ are independent of x (see Figure 4). The resulting matrix $\mathbf{X}_S$ is shown in Figure 5.

APPENDIX B: THE NUMBERS OF NUCLEOTIDE SUBSTITUTIONS OCCURRING IN TWO INTERACTING SITES IN A STEM FOLLOW TWO INDEPENDENT DISTRIBUTIONS

One can verify this assertion in the following way. Consider the kth pair of sites in the stem region that evolve with the relative rate X_k. Denote by n_1 and n_2 the numbers of nucleotide substitutions that occurred in sites 1 and 2 within the kth pair, respectively. As far as the process of dinucleotide substitution is Markovian, the actual number of transitions between 16 states for a fixed time interval follows a Poisson distribution. Because only dinucleotide states one difference apart are adjacent (see Figure 3), the total number of transitions between states is equal to the total number of nucleotide substitutions in both sites, $n_1 + n_2$. Therefore, the probability of observing $i + j$ substitutions in two sites is

$$P\{n_1 + n_2 = i + j\} = \frac{\mu^{(i+j)}e^{-\mu}}{(i + j)!}, \quad \text{(B1)}$$

where μ is the expected number of transitions between dinucleotide states for amount of time t, where $\mu = [48(\pi_p + \pi_u)\beta X_k t]$. Due to the symmetry of the adjacency graph (Figure 3) the probability of observing a substitution in site 1 is always exactly equal to probability of observing substitution in site 2 for any dinucleotide state. Therefore, the conditional probability of having i substitutions in site 1 given the total number of substitutions in both sites is $(i + j)$ is equal to

$$P\{n_1 = i \mid n_1 + n_2 = i + j\} = 2^{-(i+j)}\binom{i + j}{i}. \quad \text{(B2)}$$

Combining (B1) and (B2), we obtain the joint distribution of n_1 and n_2.

$$P\{n_1 = i, n_2 = j\} = P\{n_1 = i \mid n_1 + n_2 = i + j\}P\{n_1 + n_2 = i + j\} = \frac{(\mu/2)^i e^{-\mu/2}}{i!} \times \frac{(\mu/2)^j e^{-\mu/2}}{j!}. \quad \text{(B3)}$$

Therefore, n_1 and n_2 are independent variables following the same Poisson distribution with parameter $\mu/2$.