A Space-Time Process Model for the Evolution of DNA Sequences

Ziheng Yang

Department of Zoology, The Natural History Museum, London SW7 5BD, United Kingdom and College of Animal Science and Technology, Beijing Agricultural University, Beijing 100094, China

Manuscript received April 29, 1994
Accepted for publication October 8, 1994

ABSTRACT

We describe a model for the evolution of DNA sequences by nucleotide substitution, whereby nucleotide sites in the sequence evolve over time, whereas the rates of substitution are variable and correlated over sites. The temporal process used to describe substitutions between nucleotides is a continuous-time Markov process, with the four nucleotides as the states. The spatial process used to describe variation and dependence of substitution rates over sites is based on a serially correlated gamma distribution, *i.e.*, an *auto-gamma* model assuming Markov-dependence of rates at adjacent sites. To achieve computational efficiency, we use several equal-probability categories to approximate the gamma distribution, and the result is an *auto-discrete-gamma* model for rates over sites. Correlation of rates at sites then is modeled by the Markov chain transition of rates at adjacent sites from one rate category to another, the states of the chain being the rate categories. Two versions of nonparametric models, which place no restrictions on the distributional forms of rates for sites, also are considered, assuming either independence or Markov dependence. The models are applied to data of a segment of mitochondrial genome from nine primate species. Model parameters are estimated by the maximum likelihood method, and models are compared by the likelihood ratio test. Tremendous variation of rates among sites in the sequence is revealed by the analyses, and when rate differences for different codon positions are appropriately accounted for in the models, substitution rates at adjacent sites are found to be strongly (positively) correlated. Robustness of the results to uncertainty of the phylogenetic tree linking the species is examined.

COMPARISON of homologous DNA sequences from living species has provided an important tool for studying molecular sequence evolution. FELSENSTEIN (1981) described a maximum likelihood framework for modeling the process of nucleotide substitution combined with phylogenetic tree estimation. The model suggested by FELSENSTEIN assumes constant rate of substitution among nucleotide sites. This assumption has long been recognized as unrealistic, especially for genes that code for proteins or sequences that are otherwise functional (see WAKELEY 1993 and references therein). The most important reason appears to be that different sites perform different structural and functional roles in the gene and are therefore under different selective constraints; this leads to variable rates of substitution at sites. Mutation rates may also be variable in different regions of the genome (WOLFE *et al.* 1989).

There have been many attempts to account for rate variation among sites in nucleotide-substitution models. For example, JIN and NEI (1990) and TAMURA and NEI (1993) used the gamma distribution with given parameters to describe variable rates at sites when they constructed formulae for estimating the distance between two homologous DNA sequences. The gamma-distribution model has also been extended to a joint likelihood analysis of all sequences by YANG (1993), which is a direct extension of FELSENSTEIN's (1981) model of a single rate for all sites. Unfortunately the computation required by this method is very intensive, and YANG (1994) suggested the use of a discrete distribution as an approximation to the (continuous) gamma. Use of the gamma distribution to describe rate variation among sites has been found to produce quite good fit to various data sets (see, *e.g.*, WAKELEY 1993; YANG 1994; YANG *et al.* 1994).

The existence of "conservative" and "variable" regions in a gene suggests that rates of substitution may be not only variable, but also correlated, as sites within the same region may have similar rates characterized by the structural and functional importance of the whole region. In this paper, we will develop models that allow for such correlation by assuming Markov dependence of rates at adjacent sites. Such models will provide an alternative hypothesis for testing rate constancy and independence over sites, will produce more accurate prediction of rates at sites and will be useful for studying the effects of rate variation and correlation on various aspects of phylogenetic analysis.

The resulting models are space-time process models, by which the nucleotides in the sequence evolve over time, whereas the rates of change are variable and de-

Address for correspondence: Institute of Molecular Evolutionary Genetics, 328 Mueller Lab, University Park, PA 16802.

pendent over sites and are characterized by a spatial process. Our emphasis is on the spatial process used to model variation and dependence of rates over sites, but the temporal process of nucleotide substitution will first be described to introduce the necessary notation. The models will be applied to data of a segment of the mitochondrial genome from several primate species. We emphasize comparison of models, estimation of parameters and prediction of rates for sites as means for understanding the mechanisms of molecular sequence evolution.

THEORY

We consider substitutions only and ignore insertions and deletions. The data consist of S homologous DNA sequences from living species, each of N nucleotides long, and can be represented by an $S \times N$ matrix, $\mathbf{X} = \{x_{sn}\}$, where x_{sn} means the nth nucleotide in the sth sequence; x_{sn} takes a value from 1, 2, 3 or 4, representing the four nucleotides, T, C, A or G, respectively. We use $\mathbf{x}_n$ to denote one column in $\mathbf{X}$, which is the nucleotide composition in different species at the nth site. It is apparent that $\mathbf{x}_n$ is one of $4 \times 4 \times \cdots \times 4 = 4^S$ possible "site patterns" (see, *e.g.*, GOLDMAN 1993). The species (and their representative sequences) are related according to an evolutionary tree; an "unrooted" tree topology for four species ($S = 4$) is shown in Figure 1, which will be used as an example to develop the theory. The sequences for extinct common ancestors, *e.g.*, those at nodes 5 and 6 in the tree of Figure 1, existed in the past and are now unknown. The sequences are assumed to evolve independently of each other after the separation of the species.

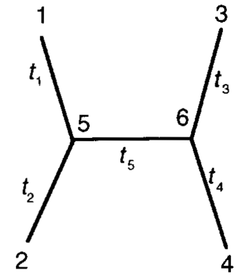

FIGURE 1.—An unrooted tree topology with four species used to develop the theory. The branch length is measured by the average number of nucleotide substitutions per site that have accumulated along the branch.

We assume that, for each site in the sequence, there is an overall rate of substitution that is determined by the structural and functional role of the site in the gene. This assumption appears legitimate when the sequences are not very different and homologous sites in different sequences perform more or less the same roles. YANG (1993, 1994) considered the case where rates for sites are variable but independent; in this paper, we extend the theory to allow for correlation of rates at adjacent sites. We assume that conditional on the rates, substitutions occur independently at different sites. This is referred to as the *conditional independence*.

The temporal process: the Markov process model of nucleotide substitution: Nucleotide substitution is assumed to follow a (stationary) homogeneous Markov process, the four nucleotides being the states of the process. Let $\mathbf{Q} = \{Q_{\mu\nu}\}$ be the *rate matrix* of the process for a site with an average overall rate. We use the substitution model proposed by HASEGAWA *et al.* (1985), by which

$$\mathbf{Q} = \begin{bmatrix} -(\kappa\pi_C + \pi_A + \pi_G) & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & -(\kappa\pi_T + \pi_A + \pi_G) & \pi_A & \pi_G \\ \pi_T & \pi_C & -(\kappa\pi_G + \pi_T + \pi_C) & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & -(\kappa\pi_A + \pi_T + \pi_C) \end{bmatrix} \cdot f \qquad (1)$$

where $Q_{\mu\nu}\Delta t$ ($\mu \neq \nu$) is the probability that nucleotide μ changes into ν in a small time interval Δt. Parameters π_ν's give the equilibrium distribution of the process, with $\Sigma\, \pi_\nu = 1$, and we assume that the process is in equilibrium. Parameter κ (usually >1) allows transitional substitutions ($T \leftrightarrow C$, $A \leftrightarrow G$) to occur with higher rates than transversional substitutions ($T, C \leftrightarrow A, G$). The model will be designated "HKY." The row sums of $\mathbf{Q}$ are 0; this allows the matrix of transition probabilities in time t to be calculated as $\mathbf{P}(t) = \{P_{\mu\nu}(t)\} = \exp(t\mathbf{Q})$ (see, *e.g.*, GRIMMETT and STIRZAKER 1992, pp. 239–246). As t and $\mathbf{Q}$ occur in the form $t\mathbf{Q}$ only (in the likelihood function), we choose the scale factor $f = 1/[4\kappa(\pi_T\pi_C + \pi_A\pi_G) + 4(\pi_T + \pi_C)(\pi_A + \pi_G)]$, so that the average rate of substitution is 1 when the process is in equilibrium, *i.e.*, $-\Sigma\, \pi_\mu Q_{\mu\mu} = 1$. $\mathbf{Q}$ thus represents the *pattern* of nucleotide substitution whereas the overall amount of evolution is reflected in t. Time t, or the branch length in a tree, is then measured by the expected number of nucleotide substitutions per site that have occurred during the time interval or along the branch. We do not assume the constancy of substitution rates among lineages, an assumption known as the *molecular clock;* as a result, the placement of the root in a tree will not affect the likelihood; that is, only unrooted tree topologies can be identified (FELSENSTEIN 1981).

To calculate $\mathbf{P}(t) = \exp(t\mathbf{Q})$, we perform the spectral decomposition (diagonalization) of $\mathbf{Q}$; if $\mathbf{Q} = \mathbf{U} \cdot \text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} \cdot \mathbf{U}^{-1}$, then $\mathbf{P}(t) = \mathbf{U} \cdot \text{diag}\{\exp(\lambda_1 t), \exp(\lambda_2 t), \exp(\lambda_3 t), \exp(\lambda_4 t)\} \cdot$

$\mathbf{U}^{-1}$, where the λs are the eigenvalues of $\mathbf{Q}$ and columns of $\mathbf{U}$ are the corresponding (right) eigenvectors; those for the HKY model are given by HASEGAWA *et al.* (1985).

The overall rates for sites are assumed to be random variables, either independent or Markov dependent, as will be described later. If the rate for site n is r_n ($n = 1, 2, \ldots, N$), the rate matrix for the site will be $r_n\mathbf{Q}$, and the matrix of transition probabilities for the site will be $\mathbf{P}(r_nt) = \exp(r_nt\mathbf{Q})$. Suppose that the nucleotide composition for this site is $\mathbf{x}_n = \{x_1, x_2, x_3, x_4\}'$ (we have written $x_{1n}, x_{2n}, \cdots$ as $x_1, x_2, \cdots$ for conciseness). The conditional probability of observing $\mathbf{x}_n$, given the rate for the site r_n, is (YANG 1993)

$$f(\mathbf{x}_n|r_n) = \sum_{x_5=1}^{4} \sum_{x_6=1}^{4} \pi_{x_5} P_{x_5x_1}(r_nt_1) P_{x_5x_2}(r_nt_2) \times P_{x_5x_6}(r_nt_5) P_{x_6x_3}(r_nt_3) P_{x_6x_4}(r_nt_4). \quad (2)$$

The "root" of the tree, *i.e.*, the starting point for calculation, is (arbitrarily) fixed at node 5 in the tree of Figure 1, and π_{x_5} is the probability of observing nucleotide x_5 at node 5, given by the equilibrium distribution of the process. The summations are taken over the unknown nucleotide states (x_5 and x_6) in the extinct ancestors at nodes 5 and 6. For an arbitrary tree topology of many species, this conditional probability can be efficiently calculated using the postorder tree-traversal algorithm of FELSENSTEIN (1981).

The spatial process: Markov chain transition in the auto-discrete-gamma model of rates over sites: The gamma distribution with parameters α and β has mean α/β and variance α/β^2. Since the rate for site (r) is seen to be a scale factor, we set $\beta = \alpha$ so that $\mathcal{E}(r) = 1$ (with variance $1/\alpha$). The density function of r is then

$$g(r; \alpha) = \alpha^{\alpha}\Gamma(\alpha)^{-1}e^{-\alpha r}r^{\alpha-1}, \quad r > 0, \quad \alpha > 0. \quad (3)$$

The single parameter α is reversely related to the extent of rate variation among sites. When $\alpha \geq 1$, the distribution is ∩-shaped; $\alpha \rightarrow \infty$ reduces to the model of a single rate for all sites. When $\alpha < 1$, the distribution is highly skewed and has a L-shape, which suggests that most sites have very low rates of substitution or are nearly "invariable", and yet there are a few mutational "hot spots"; the case of $\alpha = 0.5$ is shown in Figure 2. Maximum likelihood estimates of α from real data have been in the range 0.1–1.0 (YANG 1993; YANG *et al.* 1994).

Assuming independence of rates among sites, YANG (1993) presented an approach to calculating $f(\mathbf{x}_n) = \mathcal{E}[f(\mathbf{x}_n|r_n)]$ and hence the likelihood function. Because the computation required by this model is very intensive, YANG (1994) suggested a "discrete-gamma model" (dG), whereby a discrete distribution is used to approximate the (continuous) gamma. The range of r $(0, \infty)$ is separated into K categories by $K + 1$ threshold points, $r_0^* = 0, r_1^*, r_2^*, \ldots, r_K^* = \infty$, such that each category has probability $1/K$ of occurrence (Figure 2). The mean of a category is used to represent all rates in the category. We denote using $\bar{r}_i$ the mean for the ith category, which covers the interval (r_{i-1}^*, r_i^*). For given value of parameter α, the threshold points r_i^*s and the mean rates $\bar{r}_i$s can be easily calculated (YANG 1994).

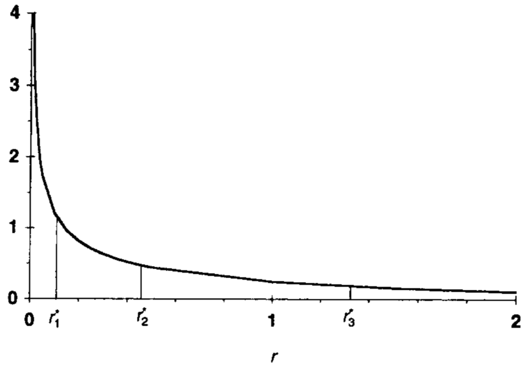

FIGURE 2.—Discretization of the (marginal) gamma distribution of rates for sites using four equal-probability categories [adapted from Figure 1 of YANG (1994)]. The distribution shown in the graph has parameter $\alpha = 1/2$, and is the χ^2 distribution with one degree of freedom. The boundaries for categories are calculated as $r_0^* = 0$, $r_1^* = 0.1015$, $r_2^* = 0.4549$, $r_3^* = 1.3233$ and $r_4^* = \infty$, which are the percentage points corresponding to $p = 0$, $1/4$, $2/4$, $3/4$ and 1. The means of the four categories are $\bar{r}_1 = 0.0334$, $\bar{r}_2 = 0.2519$, $\bar{r}_3 = 0.8203$ and $\bar{r}_4 = 2.8944$, respectively; these are used to represent all rates in each category.

In this paper, we posit Markov dependence of rates over sites. Our implementation is through this discretized gamma distribution, resulting in an *auto-discrete-gamma* model of rates for sites. We start from considering the *auto-gamma* model with rates taken as continuous random variables and then construct the discrete version as its approximation. For simplicity, we use a Markov chain to model the correlation of rates at neighboring sites; given the rate r_{n-1} for site $n - 1$, the distribution of rate r_n at site n is specified fully. It appears more realistic and natural to have r_n depend on rates at both its two neighboring sites, that is, on both r_{n-1} and r_{n+1}, which means using a *Markov random field* to model rate variation along the sequence. However, this is noted to add tremendous complexity to calculation of the likelihood function (CRESSIE 1991, pp. 383–573) and is not attempted in this study. Needless to say, we also ignore possible correlation of rates at sites separated by more than one nucleotide. We also consider an alternative model that assumes that r_n depends on only r_{n+1} instead of r_{n-1}; this turns out to give identical results for the auto-discrete-gamma model of this paper.

Consider the rates R_1 and R_2 for any two neighboring sites in the sequence, which are two (continuous) random variables. As the marginal distributions of R_1 and R_2 are both gamma, R_1 and R_2 are known to follow a *bivariate gamma distribution* (JOHNSON and KOTZ 1972, p. 216). Many such distributions have been constructed (see *e.g.*, JOHNSON and KOTZ 1972, pp. 216–230). For mathematical tractability, we have chosen to use the one due to MORAN (1969). Let Z_1 and Z_2 be two random variables with a standard bivariate normal distribution whose density is

$$(2\pi)^{-1}(1-\rho^2)^{-1/2} \times \exp\left(-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2)\right). \quad (4)$$

Define random variables U_1 and U_2 by the equations

$$U_1 = (2\pi)^{-1/2}\int_{-\infty}^{Z_1}\exp(-t^2/2)\,dt = \Phi(Z_1) \quad (5)$$

and $U_2 = \Phi(Z_2)$. The marginal distributions of U_1 and U_2 are both uniform (rectangle) in the interval (0, 1).

Now define random variables R_1 and R_2 by the equations

$$U_1 = \int_0^{R_1} g(\tau;\alpha)\,d\tau = \int_0^{R_1}\alpha^\alpha\Gamma(\alpha)^{-1}e^{-\alpha\tau}\tau^{\alpha-1}\,d\tau = G(R_1;\alpha) \quad (6)$$

and $U_2 = G(R_2;\alpha)$.

The joint probability density of R_1 and R_2 are given by MORAN (1969) for the more general case that R_1 and R_2 have marginal gamma distributions with different parameters. In our model, the spatial Markov process is assumed to be stationary, and R_1 and R_2 are assumed to have identical marginal distributions.

When we use K categories to approximate the marginal distributions of R_1 and R_2, the correlation between R_1 and R_2 will be modeled by the conditional probability that site n is from category j (with rate $\bar{r}_j$), given that site $n-1$ is from category i (with rate $\bar{r}_i$). Let Y_n be the rate category that site n is from. This probability will be $M_{ij} = \text{prob}(Y_n = j | Y_{n-1} = i) = \text{prob}(r_n = \bar{r}_j | r_{n-1} = \bar{r}_i)$. $\mathbf{M} = \{M_{ij}\}$ then constitutes the matrix of transition probabilities for a Markov chain, the states of which are the K rate categories. We calculate M_{ij} as following:

$$\begin{aligned} M_{ij} &= \text{prob}(Y_n = j | Y_{n-1} = i) \\ &\cong \text{prob}(r_{j-1}^* < R_2 < r_j^* \,|\, r_{i-1}^* < R_1 < r_i^*) \\ &= \frac{\text{prob}(r_{i-1}^* < R_1 < r_i^*,\, r_{j-1}^* < R_2 < r_j^*)}{\text{prob}(r_{i-1}^* < R_1 < r_i^*)}. \end{aligned} \quad (7)$$

Using (5) and (6), these probabilities (integrals) can be easily mapped onto the $Z_1 - Z_2$ plane. The problem turns out to be the calculation of the cumulative distribution function of a standard bivariate normal distribution, that is, $\Phi_2(z_1, z_2, \rho) = \text{prob}(Z_1 < z_1, Z_2 < z_2)$. There appears to have been much repetition and confusion in the statistics literature concerning approximate methods for calculating Φ_2. We have employed the method of OWEN (1956), based on the FORTRAN implementations of it by DONNELLY (1973) and YOUNG and MINDER (1974) (see HILL 1978; THOMAS 1979; CHOU 1985; BOYS 1989 for remarks on YOUNG and MINDER's program). The results are checked against appropriate tables published in *Biometrika* around 1930. The matrix **M** calculated in this way is symmetrical; this may be a shortcoming of the model rather than an advantage with respect to its fit to data. However, this property, together with the stationarity assumption of the Markov chain, assures that the same likelihood function is obtained no matter whether we let r_n depend on r_{n-1} only or on r_{n+1} only. The equilibrium distribution of the Markov chain specified by **M** has equal probability ($1/K$) for each rate category, in congruence with the (discretized) marginal distribution of R_1 and R_2. The model is referred to as an *auto-discrete-gamma* model of rates for sites ("AdG").

The correlation, $\rho_G = \text{corr}(R_1, R_2)$, between the two (continuous) gamma-distributed variables, is positively related to parameter ρ, which is $\rho = \text{corr}(Z_1, Z_2)$, although an algebraic relationship between the two seems difficult to obtain. The correlation (ρ_{dG}) between the rates at two neighboring sites in the auto-discrete-gamma model can be calculated as following for given values of parameters α and ρ.

$$\begin{aligned} \rho_{dG} &= \frac{\Sigma_{i=1}^K \Sigma_{j=1}^K \text{prob}(Y_{n-1} = i)\cdot M_{ij}\bar{r}_i\bar{r}_j - 1}{\Sigma_{i=1}^K \text{prob}(Y_{n-1} = i)\cdot \bar{r}_i^2 - 1} \\ &= \frac{\Sigma_{i=1}^K \Sigma_{j=1}^K \frac{1}{K}\cdot M_{ij}\bar{r}_i\bar{r}_j - 1}{\Sigma_{i=1}^K \frac{1}{K}\cdot \bar{r}_i^2 - 1} \end{aligned} \quad (8)$$

where $\text{prob}(Y_{n-1} = i) = 1/K$ according to the marginal distribution of rates for sites; the mean of the distribution is 1: $\Sigma_{i=1}^K \text{prob}(Y_{n-1} = i)\cdot\bar{r}_i = \Sigma_{i=1}^K 1/K\cdot\bar{r}_i = 1$. The relationship between parameter ρ (4) and ρ_{dG} (8) is depicted in Figure 3. When $\rho = 0$, we have $\rho_G = \rho_{dG} = 0$, $M_{ij} = \text{prob}(Y_n = j) = 1/K$, and the model reduces to the discrete-gamma model with independent rates for sites.

The likelihood function: Parameters in the auto-discrete-gamma model include $\boldsymbol{\theta} = \{\pi_T, \pi_C, \pi_A, \kappa, \alpha, \rho\}$, which are common to different tree topologies, and $\mathbf{t} = \{t_1, t_2, t_3, t_4, t_5\}$, which are branch lengths in a specific tree topology (Figure 1). Note that the joint distribution of $Y_1, Y_2, \ldots, Y_N$ is

$$\text{prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) = \text{prob}(Y_1 = y_1)M_{y_1y_2}M_{y_2y_3}\cdots M_{y_{N-1}y_N} \quad (9)$$

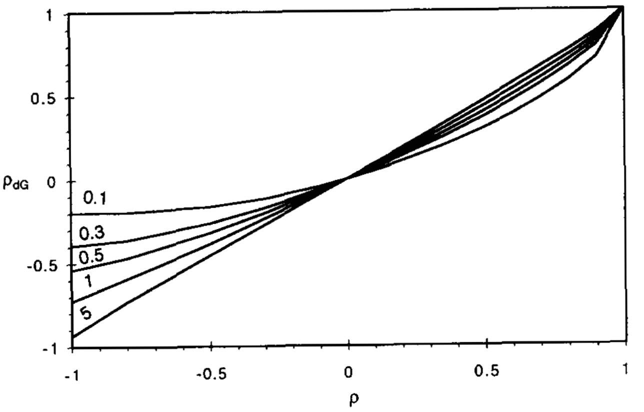

FIGURE 3.—The relationship between ρ_{dG} (8) and parameter ρ (4). The curves are for different values of the α parameter: 0.1, 0.3, 0.5, 1 and 5.

where we set prob $(Y_1 = y_1) = 1/K$, due to the stationarity of the chain. With the assumption of conditional independence of data over sites given the rates, the likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}, \mathbf{t}; \mathbf{X}) & \\ &= \operatorname{prob}(\mathbf{X}; \boldsymbol{\theta}, \mathbf{t}) \\ &= \sum_{y_1=1}^{K} \cdots \sum_{y_N=1}^{K} \Bigg(\operatorname{prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) \\ &\quad \times \prod_{n=1}^{N} f(\mathbf{x}_n | r_n = \bar{r}_{y_n}) \Bigg) \\ &= \sum_{y_1=1}^{K} \cdots \sum_{y_N=1}^{K} \Bigg(\operatorname{prob}(Y_1 = y_1) f(\mathbf{x}_1 | r_1 = \bar{r}_{y_1}) \\ &\quad \times \prod_{n=2}^{N} M_{y_{n-1}y_n} \cdot f(\mathbf{x}_n | r_n = \bar{r}_{y_n}) \Bigg) \qquad (10) \end{aligned}$$

where $f(\mathbf{x}_n | r_n = \bar{r}_{y_n})$ is given in (2). As the summation signs in (10) can be moved rightward, a simple algorithm is possible for calculating the likelihood function. Let $b_i(n) = \operatorname{prob}(\mathbf{x}_n, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | Y_n = i)$ be the probability of observing data $\mathbf{x}_n, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_N$, given that site n is from rate category i. Then

$$b_i(n) = f(\mathbf{x}_n | r_n = \bar{r}_i) \sum_{j=1}^{K} M_{ij} \cdot b_j(n+1) \qquad (11)$$

with $b_i(N) = f(\mathbf{x}_N | r_N = \bar{r}_i)$. The likelihood is simply

$$L = \sum_{i=1}^{K} \operatorname{prob}(Y_1 = i) \cdot b_i(1). \qquad (12)$$

The computation required by this model is then only slightly more than that for the discrete-gamma model assuming independent rates for sites, which is roughly K times that required by a model assuming constant rate for all sites. As a common practice, we estimate the parameters π_T, π_C and π_A in the HKY model (1) by using the averages of the observed nucleotide frequencies in the sequences. Other parameters are estimated by maximizing the likelihood function. In theory, any numerical optimization algorithms can be used for this purpose, and an EM algorithm for this type of model was described by LEROUX and PUTERMAN (1992). In this paper, a quasi-Newton algorithm is used to obtain estimates of parameters by iteration, with the gradients calculated using the difference method.

Accounting for rate differences at different codon positions: Sometimes sites in the sequence can be naturally grouped into different classes, which are known to change at different rates. This is the case for protein-coding DNA sequences, where the three codon positions are known to change at quite different rates due to the different selective constraints exerted on them; mutations at the third position may not cause changes of the amino acids whereas those at the second position always do. Another possibility is when several genes (of the same species) are combined into one data set, and different genes can be assumed to evolve with different rates determined by their relative conservativity. It seems reasonable to assign different rate parameters for sites from such different classes. If there are g site classes, we can assume that sites from class j ($j = 1, 2, \ldots, g$) have rate c_j, with $c_1 = 1$; the cs are rate ratios. We will very loosely refer to such site classes as "codon positions" and designate models that use different rates for different classes of sites as "C".

The rate matrix for a site which is from the jth site class and which has a gamma-distributed rate r is then $rc_j\mathbf{Q}$, with transition probability matrix $\mathbf{P}(t) = \exp(rc_jt\mathbf{Q})$. The likelihood function can be calculated as before, although the treatment of the cs is different from that of the rs. Simply, rates for codon positions are parameters and rates from the gamma distribution are random variables. Given any site, we know which codon position it is from and hence its rate parameter c_j. However, we do not know what value of r corresponds to the site. The likelihood function is obtained by summing over all possibilities for the random variables rs but not over the parameters cs.

Prediction of substitution rates at sites: We study the conditional distribution of rates for sites (the rs) given the data ($\mathbf{X}$). With the assumption of independent rates over sites, YANG and WANG (1994) have noted that use of the conditional mean, $\hat{r} = \mathcal{E}(r|\mathbf{x})$, as the predictor of the true rate (r) for a site with data $\mathbf{x}$ maximizes the correlation between the predictor and the true rate. Specifically, for any other predictor $\tilde{r} = \tilde{r}(\mathbf{x})$, we have $\operatorname{corr}(\tilde{r}, r) = \operatorname{corr}(\tilde{r}, \hat{r}) \cdot \operatorname{corr}(\hat{r}, r)$. For the auto-discrete-gamma model, this can similarly be defined as

$$\hat{r}_n = \mathcal{E}(r_n|\mathbf{X}) = \mathcal{E}(r_n|\mathbf{x}_n, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_N)$$

$$\cong \frac{\sum_{i=1}^{K} \bar{r}_i \cdot \text{prob}(Y_n = i) \cdot \text{prob}(\mathbf{x}_n, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | Y_n = i)}{\sum_{i=1}^{K} \text{prob}(Y_n = i) \cdot \text{prob}(\mathbf{x}_n, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | Y_n = i)}$$

$$= \frac{\sum_{i=1}^{K} \bar{r}_i \cdot \text{prob}(Y_n = i) \cdot b_i(n)}{\sum_{i=1}^{K} \text{prob}(Y_n = i) \cdot b_i(n)} \qquad (13)$$

where $\text{prob}(Y_n = i) = 1/K$, and $b_i(n)$ is defined in (11) and calculated at the maximum likelihood estimates of parameters. Alternatively, the mode of the conditional distribution may be used, which will result in the maximum *a posterior* predictor. This means using $\bar{r}_i$ as the predicted rate for site n, by which i maximizes $f(\mathbf{x}_n|r_n = \bar{r}_i)$ or $b_i(n)$ for the discrete-gamma model or auto-discrete-gamma model, respectively. However, this is found to be less, and sometimes much less, efficient than use of the conditional mean for the discrete-gamma model assuming independence of rates over sites (results not shown), presumably because the conditional distribution $(r_n|\mathbf{x}_n)$ is most often highly skewed and the mode of the discrete distribution is not very representative of the whole distribution. The mode of the *a prior* (continuous) distribution (3) does not exist for $\alpha < 1$. We expect this to be true also for Markov-dependent rates over sites, and use (13) to predict rates. Rates calculated according to (13) are normally not equal to any of the $\bar{r}_i$s ($i = 1, 2, \ldots, K$); we suggest that this is justifiable as we consider the discrete gamma model as an approximation to the continuous gamma.

Nonparametric models of rates over sites: We have also considered models for variable rates over sites, either independent or Markov-dependent, without assuming a specific distributional form for the rates. Simply, the discrete-gamma model's $\bar{r}_i$s and f_is, which used to be functions of parameter α, and the auto-discrete-gamma model's $\bar{r}_i$s and M_{ij}s, which used to be functions of parameters α and ρ, are now taken as free parameters. Let K be the number of categories of rates. Such a model includes $2(K - 1)$ free parameters when rates over sites are assumed to be independent; these are the frequencies for the rate categories: $f_1, f_2, \ldots, f_{K-1}$ (f_K is not a free parameter as $\Sigma f_i = 1$) and the rates for the categories $\bar{r}_1, \bar{r}_2, \bar{r}_{K-1}$ ($\bar{r}_K$ is given by the requirement $\Sigma f_i \bar{r}_i = 1$). The nonparametric model assuming Markov dependence will involve $(K + 1)(K - 1)$ parameters. These are the rates for the categories $\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_{K-1}$, and the $K \times K$ elements of the matrix $\mathbf{M}$ with the restriction that the K row sums of $\mathbf{M}$ are all one; the frequencies for categories (the f_is) are given by the equilibrium distribution of the Markov chain specified by $\mathbf{M}$.

Clearly the nonparametric models formulated above involve many parameters, especially when more than two rate categories are considered. We therefore consider another version of these models, with the restriction that each rate category has equal probability of occurrence. With independent rates for sites, this means $(K - 1)$ free parameters (the $\bar{r}_i$s); with Markov dependence, this restriction means that both the row sums and the column sums of $\mathbf{M}$ are one and $\mathbf{M}$ is known as a *double stochastic matrix;* the model then involves $K(K - 1)$ parameters.

Maximum likelihood estimation of parameters in these models and prediction of rates by (13) can proceed in a way similar to the auto-discrete-gamma model described before.

ANALYSIS OF PRIMATE MITOCHONDRIAL DNA (mtDNA) SEQUENCES

Data: BROWN *et al.* (1982) determined the sequences of a segment of the mitochondrial genome from human, chimpanzee, gorilla, orangutan and gibbon. There are 896 nucleotide sites in the sequences except that orangutan has a nucleotide missing at position 560. The beginning part of this segment (nucleotides 1–458) codes for part of protein ND4 (NADH-dehydrogenase subunit 4) and the ending part (nucleotides 658–896) codes for part of protein ND5 (NADH-dehydrogenase subunit 5). The middle of the segment (nucleotides 459–657) codes for three tRNAs, *i.e.*, histidine, serine and leucine tRNAs (BROWN *et al.* 1982). Sequences of this region are now also available for several other primates, and we have added those for crab-eating macaque, squirrel monkey, tarsier and lemur (HAYASAKA *et al.* 1988), so that the expanded data set contains nine species. The sequences were aligned by A. FRIDAY. Several sites in the tRNA-coding region involve gaps (insertions or deletions), and these are excluded, with 888 nucleotides left in each sequence. We note that possible errors in the alignment or the removal of sites involving gaps may bias the analysis, because consecutive sites in the resulting data may not in fact be direct neighbors, as is assumed in the models of Markov-dependent rates for sites. However, we expect such bias to be small for the current data set, as the sequences are very similar so that the alignment appears quite reliable and only a few sites in the tRNA-coding region are removed.

The 888 nucleotide sites in the data can be naturally grouped into four classes, those at the first, second, third codon positions in the two protein-coding regions and those within the tRNA-coding region. There are 233, 232, 232 and 191 sites in the four classes, respectively, and we assign rate parameters $c_1 = 1$, c_2, c_3, c_4 for them, respectively. We will call them different "codon positions". The averages of the observed nucleotide

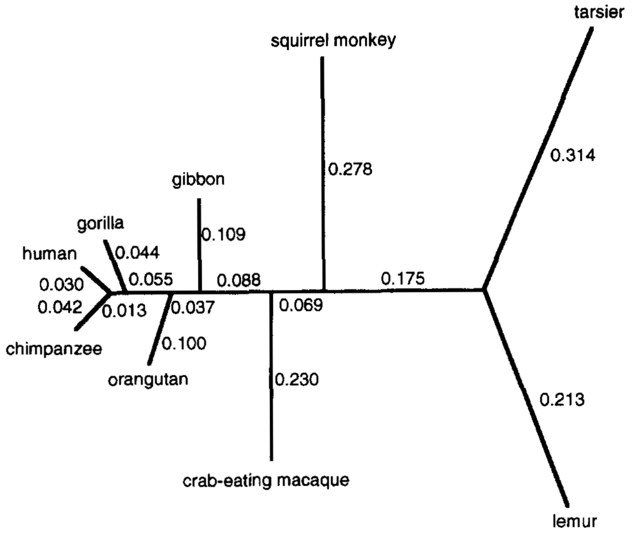

FIGURE 4.—The phylogenetic tree for nine primates whose mtDNA sequences (888 bp) are analyzed in this paper. Branch lengths shown in the tree are calculated under the HKY+C+AdG model, measured as the average numbers of substitutions per site at the first codon position. This tree topology (but not the branch lengths) is assumed to compare models and predict rates for sites in the paper.

frequencies in the whole sequences are $\hat{\pi}_T = 0.266$, $\hat{\pi}_C = 0.304$, $\hat{\pi}_A = 0.322$ and $\hat{\pi}_G = 0.108$, and these are taken as estimates of the frequency parameters in the HKY substitution model (1). Nucleotide frequencies in different species, either for the whole sequence or at different codon positions, are similar, which suggests that the temporal processes of nucleotide substitution are more or less homogeneous and stationary. However, nucleotide frequencies at different codon positions are quite different; for example, the frequencies are 0.209 (T), 0.273 (C), 0.385 (A), 0.133 (G) at the first codon position whereas they are 0.179 (T), 0.413 (C), 0.365 (A), 0.042 (G) at the third position. Our models assume one common $\mathbf{Q}$ (and thus one set of frequency parameters) for all the codon positions and are not adequate in this respect. It is possible to modify the models to allow for this feature by using different frequency parameters in the HKY model for sites from different codon positions; this is not pursued here, and we suggest that our analyses of rate variation along the sequence will not be influenced much by this inaccuracy of the models.

The phylogenetic relationship among the species may be represented as the tree shown in Figure 4. There exists controversy about the positions of tarsier and lemur (A. FRIDAY, personal communication), but this is concerned with only the placement of the root in the tree. The human-chimpanzee-gorilla separation does not seem to be very controversial anymore, and general opinion appears to support the ((human, chimpanzee), gorilla) relationship. We will use the tree topology shown in Figure 4 to estimate parameters, compare models and predict rates. The effects on these analyses due to the uncertainty of the phylogenetic relationship will be discussed later, together with the problem of estimating the phylogeny using different models.

Estimation of parameters from the parametric models: The auto-discrete-gamma model reduces to the model of a single rate for all sites when there is only one rate category ($K = 1$). When $K \to \infty$, the model will approach the (continuous) auto-gamma model. We expect that the likelihood values and parameter estimates will change dramatically for small values of K, but when K is sufficiently large, the results will stabilize. YANG (1994) analyzed several quite different data sets using the discrete-gamma model assuming independent rates over sites; different values of K were used, including the (continuous) gamma model of YANG (1993) which corresponds to $K = \infty$. Such comparisons suggest that four rate categories can provide optimum or near-optimum fit by the model to data and also quite good approximation to the (continuous) gamma distribution as reflected by the estimated α parameter. In this paper, we have introduced Markov dependence of rates into the models, but have not implemented the (continuous) auto-gamma model ($K = \infty$). Instead we perform all analyses using two values of K (4 and 8) to get some feel about the effect of K. The results, *i.e.*, likelihood values, parameter estimates and predicted rates obtained from using these two values of K turn out to be quite similar, which suggests that four categories may be sufficient for the auto-discrete-gamma model for analyzing real data, just as in the case of the discrete-gamma model (YANG 1994). In the following, we present results obtained by using $K = 8$, with comments given on those obtained from using $K = 4$.

Log-likelihood values and parameter estimates obtained under different models are shown in Table 1. The simplest model (HKY) assumes a single rate for all sites, which gives log-likelihood $l = -5234.64$ with $\hat{\kappa} = 4.217 \pm 0.292$ (standard errors are obtained by inverting the matrix of second-order derivatives of the log-likelihood with respect to parameters, calculated by the difference method). Either assuming discrete-gamma rates for sites (HKY+dG) or using different rate parameters for codon positions (HKY+C) leads to tremendous improvement in likelihood, suggesting the existence of severe rate variation among sites in the sequence. In fact, neither the discrete-gamma model nor the rates for codon positions alone can account for the rate variation observed in these data, since HKY+C+dG is significantly better than either HKY+dG (comparison of $2\Delta l = 251.62$ with χ^2_3, $P < 0.01$) or HKY+C (comparison of $2\Delta l = 131.94$ with χ^2_1, $P < 0.01$). Substitution rates for sites at the first, second and third codon positions and for sites in the tRNA-

TABLE 1

Log-likelihood values and parameter estimates under different parametric models

Model	l	$\hat{\kappa}$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$
Constant rate for all sites or within codon position ($\alpha = \infty$, $\rho = 0$)							
HKY (4)	−5234.64	4.217					
HKY+C (7)	−4983.80	5.356			0.469	3.245	0.588
Independent rates over sites ($\rho = 0$)							
HKY+dG (5)	−5043.64	8.107	0.432				
HKY+C+dG (8)	−4917.83	7.969	0.884		0.359	4.054	0.500
Markov-dependent rates over sites							
HKY+AdG (6)	−5039.10	8.029	0.430	0.168			
HKY+C+AdG (9)	−4882.35	7.858	0.865	0.623	0.361	4.353	0.634

Values in parentheses are the numbers of free parameters in the models, not including branch lengths. Parameters are estimated assuming the tree topology of Figure 4, and estimates of branch lengths are not shown. The frequency parameters in the HKY model (π_T, π_C and π_A) are estimated by using the averages of observed frequencies in the sequences. Models with dG assume (independent) discrete-gamma rates over sites, whereas those with AdG assume the auto-discrete-gamma rates; $K = 8$ rate categories are used in both cases. Models with C assume different rate parameters for codon positions: $c_1 = 1$, c_2, c_3 and c_4 for sites at the first, second and third codon positions in the protein-coding regions and for those in the tRNA-coding region, respectively.

coding region are quite different. They are in the proportion $c_1:\hat{c}_2:\hat{c}_3:\hat{c}_4 = 1:0.359:4.054:0.500$ by the HKY+C+dG model, *i.e.*, the third codon position changes >10 times faster than the second, and also sites in the tRNA-coding region change more slowly than the first codon position in the protein-coding regions. Furthermore, different sites at the same codon position also have quite different rates of substitution. The estimate of α under the HKY+C+dG model ($\hat{\alpha} = 0.884 \pm 0.124$) is larger than that under HKY+dG ($\hat{\alpha} = 0.432 \pm 0.043$); this is obviously because the rate parameters for codon positions (the cs) in the HKY+C+dG model have explained substantial part of the rate variation. However, estimates of rate parameters for codon positions (the cs) remain more or less the same whether or not the (discrete) gamma model is assumed to account for the remaining rate variation. Parameter κ and branch lengths (not shown) are severely underestimated when rate variation among sites exists but is ignored in the models, as observed by YANG *et al.* (1994); also WAKELEY 1994).

Use of the HKY+AdG model assuming Markov dependence leads to $\hat{\alpha} = 0.430 \pm 0.044$ and $\hat{\rho} = 0.168 \pm 0.056$. These values of α and ρ give $\rho_{dG} = 0.121$ by (8). Although $\hat{\rho}$ is significantly greater than 0, the serial correlation is not very strong. (The likelihood ratio test for the null hypothesis of rate independence over sites, *i.e.*, $\rho = 0$, means comparison of $2\Delta l = 9.08$ with χ_1^2, $P < 0.01$). Nevertheless, when rate differences at the codon positions are accounted for in the model (HKY+C+AdG), the estimate of ρ is much higher, *i.e.*, $\hat{\rho} = 0.623 \pm 0.060$; this value of ρ, together with $\hat{\alpha} = 0.865 \pm 0.124$, gives $\rho_{dG} = 0.544$ by (8). The increase in log-likelihood by introducing auto-correlation is also much greater; *i.e.*, $2\Delta l = 70.96$ ($P < 0.01$). These results suggest very strong correlation of rates at adjacent sites. The reason for the difference between the two estimates of ρ is that before accounting for rate differences at codon positions (HKY+AdG), rates at sites three nucleotides apart are highly correlated, so that the correlation between rates for two adjacent sites is weakened (see results concerning predicted rates for sites below).

Using four rate categories ($K = 4$) rather than eight in the above comparisons would give essentially identical results. The estimates of parameters are also very similar. For example, those obtained from the HKY+C+AdG model ($K = 4$) are $\hat{\kappa} = 7.843$, $\hat{\alpha} = 0.866$, $\hat{\rho} = 0.665$, $\hat{c}_2 = 0.361$, $\hat{c}_3 = 4.361$, $\hat{c}_4 = 0.639$, with $l = -4883.67$ (*cf.* Table 1). When $K = 1, 2, 3, 4, 8$ and 20, the log-likelihood for the HKY+C+AdG model is −4983.80, −4890.47, −4884.88, −4883.67, −4882.35 and −4881.70, respectively. Two categories can explain substantial part of the rate variation, and the results tend to stabilize when three or four categories are used.

Concerning the effects that ignoring the correlation of rates over sites (ρ) has on the estimation of other parameters, we note that estimates of κ, α and the rate parameters for codon positions (the cs) are quite stable whether independence ($\rho = 0$) or Markov dependence is assumed for rates among sites. Estimated branch lengths for the auto-discrete-gamma models are also very similar to but are all slightly smaller than those for the discrete-gamma models assuming independence (results not shown); the reason for this difference is not clear. The calculated standard errors for parameter estimates are larger for the auto-discrete-gamma models with Markov dependence than those for the discrete-

TABLE 2

Log-likelihood values and parameter estimates under the nonparametric models with two rate categories

Model	l	Rate ($\bar{r}_i$)	Transition probabilities (M_{ij})		Frequency (f_i)	$\hat{\kappa}$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$
Independent rates									
HKY (6)	−5047.89	0.150			0.602	7.447			
		2.286			0.398				
HKY+C (9)	−4910.57	0.361			0.779	7.638	0.388	5.326	0.542
		3.252			0.221				
Markov-dependent rates over sites									
HKY (7)	−5045.34	0.144	0.633	0.367	0.583	7.021			
		2.196	0.513	0.487	0.417				
HKY+C (10)	−4884.91	0.330	0.842	0.158	0.686	7.753	0.368	4.453	0.575
		2.463	0.346	0.654	0.314				

See note to Table 1.

gamma models assuming independence (results not shown). This seems to be due to two reasons. First, adding parameters (ρ in this case) to a model will normally "decrease" the accuracy of estimates of other parameters. Second, the positive correlation of rates at sites implies positive correlation of data at sites, which will cause the data to contain less information than if they are independent. In sum, ignoring correlation of rates over sites when it exists does not seem to bias estimates of other parameters too much, but the calculated standard errors in the estimates will give a wrong impression of high accuracy.

The nonparametric models: We have fitted the nonparametric models to the mtDNA data, assuming either independence or Markov dependence of rates over sites. The parameter-richness of the models has led to many problems when more than three rate categories are used; these will be discussed later. The results shown in Table 2 are obtained using two rate categories ($K = 2$). Overall the same conclusions can be drawn from these results as from those in Table 1. For example, Table 2 clearly suggests that rates of substitution are different for sites at different codon positions and for different sites from the same codon position (results for models assuming a single rate for sites are listed in Table 1). The most complex model, which assumes different rate parameters for codon positions and Markov-dependent random rates over sites, is significantly better than all the simpler models. The likelihood values are so different that we do not need to consult statistical tables to perform the tests. According to this model, rates for the first, second, third codon positions in the protein-coding regions and for sites in the tRNA-coding region are in the proportion 1:0.368: 4.453:0.575. Furthermore, the remaining rate variation can be best explained by two rates 0.330, 2.463, with stationary probabilities 0.686 and 0.314, respectively. Estimates in the **M** matrix suggest that if a site is in rate category one, the next site will have probability 0.842 of being in category one too and probability 0.158 of switching into category two, whereas if a site is in category two, the next site will remain in category two with probability 0.654 and switch to category one with probability 0.346; thus rates at neighboring sites are positively correlated. Using parameters (the cs) to account for rate differences at codon positions is seen to reduce the remaining rate variation (as indicated by the smaller differences between $\bar{r}_1$ and $\bar{r}_2$), and to considerably increase the correlation of rates at neighboring sites (as reflected by larger M_{11} and M_{22}). These results are congruent with those obtained from the auto-discrete-gamma models (Table 1). Estimates of the rates (the $\bar{r}_i$s) and frequencies (the f_is) are more or less stable whether independence or Markov dependence is assumed for rates over sites. Estimates of parameters κ and the cs are also very similar to those obtained from the corresponding discrete-gamma models (Table 1).

Results obtained from the nonparametric models under the restriction that each rate category has equal probability of occurrence are presented in Table 3. Because of this restriction, we have been able to obtain results with either 2 or 3 rate categories. The frequency for each category is $f_i = {}^1/_2$ or ${}^1/_3$ for $K = 2$ or 3, respectively. Note that models with two rate categories ($K = 2$ in Table 3) are equivalent to the corresponding (auto-) discrete-gamma models with two rate categories; $\bar{r}_1$ and M_{11} in the current models correspond to α and ρ in the auto-discrete-gamma models through a reparameterization. Again, the most complex model is much better than all the simpler ones either for $K = 2$ or $K = 3$; rates for sites at different codon positions are different, rates at sites within the same codon position are variable, and rates at adjacent sites are positively correlated. Other conclusions reached from results of

TABLE 3

Log-likelihood values and parameter estimates for the nonparametric models

Model	l	Rate ($\bar{r}_i$)	Transition probabilities (m_{ij})			$\hat{\kappa}$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$
Two equal-probability categories of rates									
Independent rates									
HKY (5)	−5052.21	0.110				6.394			
		1.890							
HKY+C (8)	−4924.87	0.254				7.664	0.370	3.804	0.518
		1.746							
Markov-dependent rates									
HKY (6)	−5047.60	0.109	0.576	0.424		6.284			
		1.891	0.424	0.576					
HKY+C (9)	−4890.47	0.253	0.791	0.209		7.712	0.370	4.150	0.644
Three equal-probability categories of rates									
Independent rates									
HKY (6)	−5044.01	0.031				8.176			
		0.408							
		2.561							
HKY+C (9)	−4914.54	0.314				7.763	0.373	4.709	0.530
		0.314							
		2.372							
Markov-dependent rates									
HKY (10)	−5024.08	0.164	0.972	0.000	0.028	7.262			
		0.191	0.000	0.484	0.516				
		2.645	0.028	0.516	0.456				
HKY+C (13)	−4879.31	0.217	0.869	0.087	0.044	7.915	0.364	4.683	0.666
		0.424	0.022	0.644	0.334				
		2.359	0.109	0.269	0.622				

Values and estimates given for models under the restriction of equal probability in each rate category using two or three rate categories.

Tables 1 or 2 are also apparent in Table 3. Estimates of κ and the cs are quite similar to those in Tables 1 and 2. The models with two or three rate categories (Table 3) are not nested so that a likelihood ratio test cannot be applied to compare them, but it seems that three rate categories are worthwhile.

The problem of phylogenetic trees: FELSENSTEIN (1981; see also FELSENSTEIN 1973; THOMPSON 1975) suggests that the likelihood values calculated for different tree topologies can be compared to estimate the phylogenetic relationship among the species. The method is known as maximum likelihood estimation of the phylogenetic tree. Estimation of phylogeny from DNA sequences has been of great interest to evolutionary biologists, and one may (rightly) require that an adequate model be used in such an adventure. This paper focuses on construction and comparison of models as means for understanding the processes of DNA sequence evolution. Strictly speaking, comparison of models, especially by using the chi-square approximation to the likelihood ratio test, requires the likelihood values to be calculated (and parameters to be estimated) using the true phylogenetic relationship (Z. YANG, N. GOLDMAN and A. FRIDAY, unpublished data). In practice, the difficulties involved in these two interrelated problems are quite different. In the following, we give a short discussion on the implications of results of this study to the two problems.

First, our ignorance or uncertainty concerning the phylogenetic relationship does not seem to introduce much error in the estimation of parameters in the evolutionary models or in the comparison of such models. There are $1 \times 3 \times \cdots \times (2 \times 9 - 5) = 135{,}135$ possible bifurcating tree topologies for nine species. To see the effects of changes to tree topologies, we have performed all analyses described above using several other tree topologies although results obtained from the tree of Figure 4 only are presented (Tables 1–3). As an example, we list in Table 4 the likelihood values and parameter estimates obtained under the HKY+C+AdG model for these tree topologies. The nine-species star tree has only nine branches. Other tree topologies used differ from the tree of Figure 4 only concerning the human-chimpanzee-gorilla separation. Let T_1 = ((HC)G) represent the tree of Figure 4, and then the other trees can be represented as T_2 =

TABLE 4

Log-likelihood values and parameter estimates under the HKY+C+AdG model for several different tree topologies

Tree	l	$\hat{\kappa}$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$
Star tree	−5016.46	15.358	0.346	0.455	0.214	5.466	0.527
T_0 = (HCG)	−4888.45	7.825	0.861	0.621	0.366	4.428	0.654
T_1 = ((HC)G)	−4882.35	7.858	0.865	0.623	0.361	4.353	0.634
T_2 = ((HG)C)	−4888.44	7.827	0.861	0.621	0.366	4.426	0.654
T_3 = ((CG)H)	−4885.96	7.660	0.893	0.628	0.372	4.456	0.648

((HG)C), T_3 = ((CG)H) and T_0 = (HCG), T_0 having a trifurcation. The star tree is quite different from T_0, T_1, T_2 or T_3, and estimates of parameters obtained for this tree are admittedly quite different from those for other trees. However, parameter estimates obtained for T_0, T_1, T_2 or T_3, which are not too wrong and may all be called "reasonable" trees, are very similar. Likelihood values for different trees are not very different, especially in comparison to the dramatic changes in likelihood due to changes in the assumed models (see Tables 1, 2 and 3). The same pattern is observed for other models considered in this paper (see YANG *et al.* 1994 for more examples). This means that we will obtain essentially the same results concerning parameter estimation and model selection, no matter which of the reasonable trees is used.

In contrast, the small differences in likelihood among tree topologies suggest the difficulty of phylogenetic tree estimation; some of the theoretical difficulties are discussed by Z. YANG, N. GOLDMAN and A. FRIDAY (unpublished data). It has been observed that ignoring rate variation over sites can substantially influence phylogenetic tree reconstruction, especially the estimation of branch lengths (YANG *et al.* 1994; also WAKELEY 1994). Nevertheless, this study tends to suggest that ignoring the correlation of rates over sites will not influence phylogenetic tree reconstruction greatly, at least if the point estimation only is concerned. For all models considered in this paper, the order of the likelihood values for the examined trees has been $l_{T_1} > l_{T_3} > l_{T_2} > l_{T_0}$; it seems very likely that T_1 (Figure 4) is the maximum likelihood tree by these models if all tree topologies could be evaluated. We suggest that for the estimation of tree topology, the discrete-gamma model is elaborate enough, and the auto-discrete-gamma model may not be worthwhile.

Prediction of rates at sites: We calculated the rates for the 888 sites in the mtDNA sequence using (13) based on maximum likelihood estimates of parameters in the models. As another way to look at rate dependence over sites, we calculated the serial correlations using the predicted rates and the results are shown in Figure 5. The correlation (0.562) of predicted rates at two adjacent sites calculated from the HKY+C+AdG model (K = 8) agrees well with ρ_{dG} = 0.544 calculated from (8) using the maximum likelihood estimates of the parameters, $\hat{\alpha}$ = 0.865 and $\hat{\rho}$ = 0.623 (Table 1). The decrease of the serial correlation with the number of nucleotides that separate the sites also agrees nicely with the model's expectation. The predicted rates can be plotted along the sequence after some smoothing and appear very useful for identifying conservative and variable regions in the sequence (results not shown).

The period of three in the curves for the HKY+dG and HKY+AdG models is clearly due to these models' failure to account for rate differences at the codon positions. In this regard, the "detrending" or removal of the large scale variation by using rate parameters for codon positions in the HKY+C+dG and HKY+C+AdG models is seen to be quite successful. We also note that the serial correlations, especially those for sites that are separated by one or two nucleotides, calculated from the HKY+C+dG model, which assumes independent rates over sites, are smaller than those obtained from

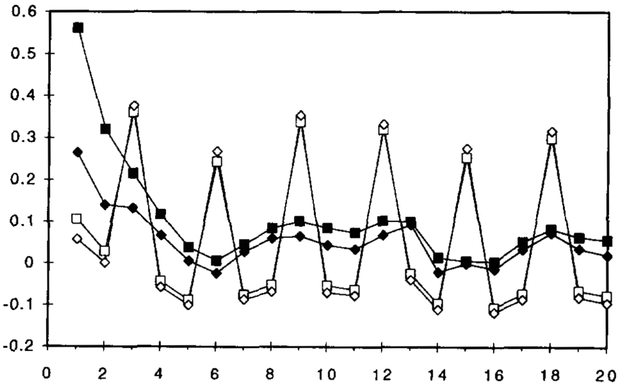

FIGURE 5.—Serial correlation of substitution rates along the mtDNA sequence, which are predicted by assuming the HKY+C+AdG (■), HKY+C+dG (◆), HKY+AdG (□) and HKY+dG (◇) models; K = 8 categories are used in these discrete-gamma models. The tree topology of Figure 4 is assumed. The graph shows the correlation coefficients between predicted rates ($\hat{r}$) at sites separated by 1, 2, . . . , 20 nucleotides. For the HKY+C+AdG and HKY+C+dG models, which assume different rate parameters for codon positions, only the random variable (r) from the (discrete) gamma distribution is used in the calculation.

TABLE 5

Correlations between rates predicted from different (auto-)discrete-gamma models using eight or four categories

	AdG+C (8)	dG+C (8)	AdG (8)	dG (8)	AdG+C (4)	dG+C (4)	AdG (4)
$K = 8$							
AdG+C							
dG+C	0.9419						
AdG	0.7773	0.8143					
dG	0.7526	0.8018	0.9983				
$K = 4$							
AdG+C	0.9939	0.9353	0.7765	0.7525			
dG+C	0.9365	0.9942	0.8163	0.8045	0.9394		
AdG	0.7720	0.8082	0.9869	0.9857	0.7763	0.8160	
dG	0.7509	0.7980	0.9859	0.9873	0.7548	0.8055	0.9987

the HKY+C+AdG model, which assumes Markov dependence. The results from the HKY+C+AdG model are clearly more reliable, and suggest that if one (wrongly) assumes rate independence over sites in the model, one will underestimate the extent of dependence, which is not very surprising.

Table 5 lists the correlations between rates predicted using different methods. If we consider the HKY+C+AdG model as giving the best predicted rates for sites, these correlations will indicate the relative efficiency of rate prediction by other models. Rates predicted using four categories have correlations ~0.99 with those using eight categories. Combined with the similarity of likelihood values and parameter estimates for these two values of K, we suggest that four categories are sufficient for analyzing real data. Rates predicted from models assuming independence (the dG models) are closely related to rates predicted from corresponding models assuming Markov-dependence (the AdG models). We also note that using one of the tree topologies T_2, T_3 or T_0 instead of T_1 produces very similar predicted rates (results not shown). Similar to the results of YANG and WANG (1994), possible errors in estimates of parameters or tree topologies normally do not affect the accuracy of rate prediction much.

DISCUSSION

The spatial-process models considered in this paper have many counterparts in various fields of applied statistics, especially in analyses of time series or spatial data. In time-series analysis, the counts of events that occur in fixed time intervals have a Poisson distribution (with the variance equal to the mean) when the process is generated by a constant homogeneous rate. When the underlying rate is variable, it is known as an overdispersed process, since the variance of the counts is larger than the mean. When the rate is itself an independent gamma variable, the counts are known to follow a negative-binomial distribution. The nonparametric models considered in this paper are known as *finite-mixture* models as the data are generated from a mixture of categories of rates with different probabilities. With Markov dependence, the models are also known as *hidden-Markov-chain* models, as the states of the chain are random variables and are not observable. LEROUX and PUTERMAN (1992) summarized recent developments of techniques concerned with the finite-mixture models. CHURCHILL (1989) employed a hidden-Markov-chain model to describe the occurrences of nucleotides in a single DNA sequence. The distinction made in this paper between the rate parameters for codon positions (the cs) and the gamma-distributed random rates (the rs) is analogous to the linear-mixed-models theory, which is widely used in animal breeding (HENDERSON 1973), although the current models are highly nonlinear; rates for codon positions are *fixed effects,* for which we estimate their *main effects* (the cs), whereas rates from the gamma distribution are *random effects,* for which we estimate their *variance components* (parameters α and ρ) and *predict* rates (the rs) based on the observed data.

Although conceptually very simple, the nonparametric models are found to be very difficult to implement in the form of a computer program, and these difficulties make them much less attractive than the parametric auto-discrete-gamma models. Suppose that a continuous distribution does provide an adequate description of the rate variation over sites; then there will be infinitely many possibilities to approximate the continuous distribution using several categories. These possibilities correspond to different values of parameters in the nonparametric models, which can be expected to have very similar likelihood values in real data analysis. Due to the near-flatness of the likelihood surface, convergence in the iteration algorithm is difficult to achieve and parameter estimates are unstable or have values at the preset boundaries such as 0 and their interpretation can be difficult. The difficulties have been encountered by LEROUX and PUTERMAN (1992) in their analysis of a sequence of counts of movements by a fetal lamb. For example, estimates obtained by those authors from a Markov-dependent mixture model with four rate categories suggest that there exists an absorbing state in the Markov chain with rate 0, which means that as soon as the fetus enters this rate category, it will never move again; because the chain is assumed to be stationary, this also means that the fetus has been and will remain motionless. We have been able to obtain equally absurd results using our nonparametric models with three or four rate categories.

The auto-discrete-gamma model (HKY+C+AdG in Table 1) and the two versions of nonparametric models assuming Markov dependence (Tables 2 and 3) are not nested, and so the likelihood ratio test is not directly applicable for comparing them. However, the likelihood values suggest that the auto-discrete-gamma model provides a better fit to the data than the nonparametric models using two rate categories. When three categories are used, the nonparametric models (*e.g.*, the last model listed in Table 3) can fit the data slightly better than the auto-discrete-gamma model, but at the cost of many more parameters. It is also noteworthy that results obtained using two or three categories in the nonparametric models are not easily comparable, but K is not an important factor in the auto-discrete-gamma model as long as a relatively large value (such as four) is used. We conclude that the auto-discrete-gamma model provides the most-parsimonious explanation of rate variation at sites in these mtDNA sequences.

C source codes are available from the author which implement models described in this paper.

I thank CLIVE MONCRIEFF for many useful discussions about the bivariate gamma distributions. This study was partially supported by a grant from the Natural Science Foundation of China.

LITERATURE CITED

BOYS, R. J., 1989 Remark AS R76: a remark on algorithm AS 76: an integral useful in calculating non-central t and bivariate normal probabilities. Appl. Statist. **38:** 580–582.

BROWN, W. M., E. M. PRAGER, A. WANG and A. C. WILSON, 1982 Mitochondrial DNA sequences of primates, tempo and mode of evolution. J. Mol. Evol. **18:** 225–239.

CHOU, Y.-M., 1985 Remark AS R55: a remark on algorithm AS 76: an integral useful in calculating non-central t and bivariate normal probabilities. Appl. Statist. **34:** 100–101.

CHURCHILL, G. A., 1989 Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. **51:** 79–94.

CRESSIE, N. A. C., 1991 *Statistics for Spatial Data.* John Wiley and Sons, New York.

DONNELLY, T. G., 1973 Algorithm 462: bivariate normal distributions. Comm. ACM **16:** 638.

FELSENSTEIN, J., 1973 Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. **22:** 240–249.

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

GOLDMAN, N., 1993 Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. **37:** 650–661.

GRIMMETT, G. R., and D. R. STIRZAKER, 1992 *Probability and Random Processes,* Ed. 5, Clarendon Press, Oxford.

HAYASAKA, K., T. GOJOBORI, and S. HORAI, 1988 Molecular phylogeny and evolution of primate mitochondrial-DNA. Mol. Biol. Evol. **5:** 626–644.

HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

HENDERSON, R., 1973 Sire evaluation and genetic trends, pp. 10–41 in *Animal Breeding and Genetics Symposium in Honour of Dr. J. L. Lush.* American Society of Animal Science and Animal Dairy Science Association, Champaign, IL.

HILL, I. D., 1978 Remark AS R26: a remark on algorithm AS 76: an integral useful in calculating non-central t and bivariate normal probabilities. Appl. Statist. **27:** 379.

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

JOHNSON, N. L., and S. KOTZ, 1972 *Statistical Distributions: Multivariate Continuous Distributions.* John Wiley and Sons, New York.

LEROUX, B. G., and M. L. PUTERMAN, 1992 Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics **48:** 545–558.

MORAN, P. A. P., 1969 Statistical inference with bivariate gamma distributions. Biometrika **56:** 627–634.

OWEN, D. B., 1956 Tables for computing bivariate normal probabilities. Ann. Math. Statist. **27:** 1075–1090.

TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

THOMAS, G. E., 1979 Remark AS R30: a remark on algorithm AS 76: an integral useful in calculating non-central t and bivariate normal probabilities. Appl. Statist. **28:** 113.

THOMPSON, E. A., 1975 *Human Evolutionary Trees.* Cambridge University Press, Cambridge.

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

WAKELEY, J., 1994 Rate variation among sites and substitutional bias among nucleotides are conflated in simple sequence comparisons. Mol. Biol. Evol. **11:** 436–442.

WOLFE, K. H., P. M. SHARP and W.-H. LI, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

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.

YANG, Z., and T. WANG, 1994 Mixed model analysis of DNA sequence evolution. Biometrics (in press).

YANG, Z., N. GOLDMAN and A. E. FRIDAY, 1994 Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol. Biol. Evol. **11:** 316–324.

YOUNG, J. C., and CH. E. MINDER, 1974 Algorithm AS 76: an integral useful in calculating non-central t and bivariate normal probabilities. Appl. Statist. **23:** 455–457. [Correction: Appl. Statist. **28:** 113 (1979)]

Communicating editor: M. BULMER

Reviewer's comments on Z. Yang's paper

'A space-time process model for the evolution of DNA sequences'

This manuscript addresses an important problem long overlooked in molecular phylogenetics. In the previous papers, Dr Yang took account of rate variation among sites in the maximum likelihood analyses. In this manuscript he extended his previous works, and took account of correlation of rates over sites. Furthermore, he applied his method to a real data of mitochondrial DNA.

His work is based on a solid statistical background, and his achievement is important to molecular phylogenetics. For this reason, I recommend publication of this paper in *Genetics.*

However, I have some minor comments.

1. page 17: 'protein 4' and 'protein 5' should be 'ND4' and 'ND5', abbreviations of NADH-dehydrogenase subunits 4 and 5.

2. page 17: No reference is given to the sequences for crab-eating macaque, squirrel monkey, tarsier and lemur. Presumably, he used data of Hayasaka et al. (Mol. Biol. Evol. 5:626–644 (1988)). The reference should be given.

Oxford
August 2, 1994

Dear Dr Yang,

I enclose two reports on your MS submitted to Genetics. I shall be glad to recommend acceptance of a revision incorporating your response to the reviewers' comments.

Please send me the revised MS at my address in New Jersey. Please enclose the reviewers' comments, since I haven't kept a copy of them.

Yours sincerely
Michael Bulmer

Comments for the author:

The manuscript by Yang addresses some statistical issues in molecular evolution regarding rate heterogeneity. A correlation component between adjacent bases is introduced as a generalization of the gamma–distributed rates model. As a general comment, the statistical aspects of this paper are addressed thoughtfully and with a precision that is rare in this type of applied work. The manuscript is well written, although the organization of later sections seems to wander a bit in its focus. Given the negative conclusion (page 25, bottom) and the realtive complexity of the methods, the practical utility of these results is not well demonstrated.

Specific comments:

1. The model of (Markov) dependence between rates at adjacent sites is motivated by the observation of "variable" and "conserved" regions in aligned sequence sets. The choice of a gamma model is arbitrary but acceptable in light of the general robustness of hierarchical models. The particular choice for the form of the bivariate model is not well justified ("by trial and error", page 9). What are the considerations behind this choice?
2. The model description which follows is (perhaps unneccesarily) complex. Could the description be simplified?
3. Is the likelihood (page 12, eqn. 10) really computable in this form? It seems to involve K^N terms.
4. Maximization of the likelihood is achieved by direct search. Does this mean grid search, Newton–Raphson or some other numerical hill–climber? Could other iterative procedures be considered here. Perhaps a variation on Felsenstein's (1981) EM algorithm could be implemented.
5. I am not convinced of the advantages of the parametric model over the non–parametric model. From an estimation point of view, the non–parametric approach clearly has some problems. But for hypothesis testing and for estimation of branch length parameters it may deserve further consideration.
6. There are some typographical errors that a good spell checker will correct.