Models of Amino Acid Substitution and Applications to Mitochondrial Protein Evolution

Ziheng Yang,† Rasmus Nielsen,‡ and Masami Hasegawa†*

*Department of Biology (Galton Laboratory), University College London; †The Institute of Statistical Mathematics, Tokyo, Japan; and ‡Department of Integrative Biology, University of California at Berkeley

Models of amino acid substitution were developed and compared using maximum likelihood. Two kinds of models are considered. ''Empirical'' models do not explicitly consider factors that shape protein evolution, but attempt to summarize the substitution pattern from large quantities of real data. ''Mechanistic'' models are formulated at the codon level and separate mutational biases at the nucleotide level from selective constraints at the amino acid level. They account for features of sequence evolution, such as transition-transversion bias and base or codon frequency biases, and make use of physicochemical distances between amino acids to specify nonsynonymous substitution rates. A general approach is presented that transforms a Markov model of codon substitution into a model of amino acid replacement. Protein sequences from the entire mitochondrial genomes of 20 mammalian species were analyzed using different models. The mechanistic models were found to fit the data better than empirical models derived from large databases. Both the mutational distance between amino acids (determined by the genetic code and mutational biases such as the transition-transversion bias) and the physicochemical distance are found to have strong effects on amino acid substitution rates. A significant proportion of amino acid substitutions appeared to have involved more than one codon position, indicating that nucleotide substitutions at neighboring sites may be correlated. Rates of amino acid substitution were found to be highly variable among sites.

Introduction

Amino acid sequences of conservative proteins are widely used to infer distant phylogenetic relationships such as early divergences near the root of the universal tree of life (Iwabe et al. 1989; Brown and Doolittle 1995; Hashimoto and Hasegawa 1996). For distant relationships, the use of nucleic acid sequences can be problematic, as alignment may be difficult, base frequencies may vary among species, and saturation of substitutions may have diluted phylogenetic information (see, e.g., Yang and Roberts 1995). In such cases, use of protein sequences may be advantageous (Hasegawa and Hashimoto 1993).

The first phylogenetic analysis of protein sequence data by a rigorous likelihood approach appears to be that of Bishop and Friday (1985, 1987). Their analysis was based on a Poisson process model of amino acid substitution assuming an equal substitution rate between any two amino acids and assuming a molecular clock (i.e., constancy of substitution rates among lineages). Neither assumption appears realistic. Kishino, Miyata, and Hasegawa (1990) relaxed these assumptions by adapting the maximum-likelihood approach of Felsenstein (1981) for nucleotide sequences, which does not rely on the existence of a molecular clock. Furthermore, the empirical matrix of substitution frequencies between amino acids compiled by Dayhoff, Schwartz, and Orcutt (1978) was used to account for different substitution rates between amino acids. The development of a computer program for protein maximum likelihood has popularized the approach (Adachi and Hasegawa 1992, 1996*a*), which is now widely used for phylogenetic analysis of protein sequence data.

The empirical model of Dayhoff, Schwartz, and Orcutt (1978) and its update by Jones, Taylor, and Thornton (1992) were constructed by averaging over many proteins over different timescales. The relative substitution rates between amino acids are fixed in those models, no matter which protein is analyzed (see Wilbur [1985], Goldman and Yang [1994], and Thorne, Goldman, and Jones [1996] for criticisms of the empirical models). However, we expect that proteins with different functions or from different genomes should have different patterns of amino acid substitution. It is thus interesting to develop models which account for the biological processes involved in amino acid substitution, i.e., mutational biases in the DNA, translation of the DNA into protein according to the genetic code, and acceptance or rejection of the resulting amino acid under selective constraints on the protein. Such ''mechanistic'' models should naturally be formulated at the codon level and may involve parameters that characterize the substitution pattern in the protein. In contrast to analysis of nucleotide sequences, for which a number of probabilistic models have been suggested (see, e.g., Yang [1994*a*] and Zharkikh [1994] for reviews), not many models of amino acid substitution are available. Several Markov-process models of codon or amino acid substitution were proposed in the literature, but they were used to predict amino acid frequencies in a protein (Jorré and Curnow 1975) or to calculate mutational distances between amino acids determined by the genetic code (Coates and Stone 1981); none of them was used in comparative analysis of real sequence data.

In this paper, codon-based mechanistic models of amino acid substitution are developed. A general approach is presented which transforms a Markov-process model of codon substitution into an amino acid substitution model by grouping synonymous codons that en-

Key words: amino acid substitution, mammals, maximum likelihood, mitochondrial DNA, protein evolution.

Address for correspondence and reprints: Ziheng Yang, Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, England. E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 15(12):1600–1611. 1998

code the same amino acid. Empirical models of amino acid substitution derived previously (Dayhoff, Schwartz, and Orcutt 1978; Jones, Taylor, and Thornton 1992) and in this paper are used for comparison. The problem of substitution rate variation among amino acid sites is also examined using the gamma-rates model of Yang (1994*b*), developed for nucleotide sequences. A large data set containing all proteins in the mitochondrial genome from 20 species of mammals was analyzed to compare the different models.

Data

The data are from Cao et al. (1998) and consist of all 12 proteins encoded by the same strand of the mitochondrial genome from 17 eutherian species and their outgroups (two marsupials and one monotreme). The 12 proteins were concatenated into one long sequence and analyzed as one data set, since they appear to have similar substitution patterns and since some of the models considered in this paper involve many parameters that require a large amount of data to estimate reliably. The other protein, ND6, is encoded by the opposite strand of the DNA with quite different base and codon biases and was not used. The species are human (*Homo sapiens,* D38112), common chimpanzee (*Pan troglodytes,* D38113), bonobo (*Pan paniscus,* D38116), gorilla (*Gorilla gorilla,* D38114), Bornean orangutan (*Pongo pygmaeus pygmaeus,* D38115), Sumatran orangutan (*Pongo pygmaeus abelii,* X97707), common gibbon (*Hylobates lar,* X99256), harbor seal (*Phoca vitulina,* X63726), grey seal (*Halichoerus grypus,* X72004), cat (*Felis catus,* UU20753), horse (*Equus caballus,* X79547), Indian rhinoceros (*Rhinoceros unicornis,* X97336), cow (*Bos taurus,* J01394), fin whale (*Balaenoptera physalus,* X61145), blue whale (*Balaenoptera musculus,* X72204), rat (*Rattus norvegicus,* X14848), mouse (*Mus musculus,* J01420), wallaroo (*Macropus robustus,* Y10524), opossum (*Didelphis virginiana,* Z29573), and platypus (*Ornithorhynchus anatinus,* X83427). After removal of sites with alignment gaps, regions in which the alignment was ambiguous, and overlapping regions between ATP6 and ATP8 and between ND4 and ND4L, the sequence had 3,331 amino acid sites. Uncertainties exist concerning the relationship among primates (A in fig. 1), ferungulates (B), rodents (C), and the outgroup (D). The most likely relationship is ((AB)CD), shown in figure 1 (Cao et al. 1998). The other two tree topologies concerning the relationship of those four groups, i.e., ((AC)BD) and ((AD)BC), were also used in fitting several models implemented in this paper. For estimating parameters and comparing models, the three tree topologies produced virtually identical results. The tree topology of figure 1 was used to obtain results presented in this paper.

Codon-based models were used to analyze both the nucleotide and amino acid sequences. For this purpose, only data from the seven primate species were used, as the nucleotide sequences from the large data set appear to be too divergent. For example, the proportions of synonymous differences between the placentals and the marsupials were often around 75%, so the method of Nei and Gojobori (1986) was either inapplicable or gave very large estimates of synonymous rates ($d_S > 2$ substitutions per site). Nonsynonymous rates in those comparisons were $d_N \approx 0.2$. Within primates, the largest estimates of synonymous and nonsynonymous rates were $d_S \approx 0.9$ and $d_N \approx 0.06$. The subset of the sequence data is referred to as the small data set.

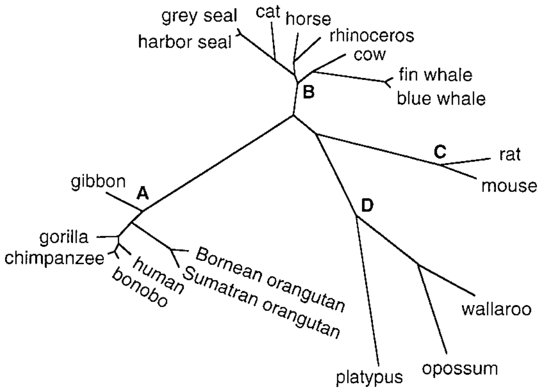

FIG. 1.—The most likely phylogeny of the 20 species analyzed in this paper (Cao et al. 1998). Branches are drawn in proportion to estimates of their lengths under the REV + G model. The tree topology (but not the branch lengths) is used in the paper to compare different amino acid substitution models.

Models of Amino Acid Substitution

Models considered in this paper differ in their assumptions concerning the pattern of amino acid substitution, i.e., the relative substitution rates between amino acids. A continuous-time Markov process is used to model amino acid substitution with a 20 × 20 matrix of instantaneous substitution rates given by $Q = \{q_{ij}\}$. The diagonals of the matrix are determined by the mathematical requirement that row sums of the matrix are all zero; that is, $q_{ii} = -\Sigma_{j\neq i}\, q_{ij}$. As time and rate are confounded in such an analysis, the rate matrix is scaled so that the average rate of substitution is one:

$$-\sum_i \pi_i q_{ii} = 1, \tag{1}$$

where π_i is the equilibrium frequency of amino acid *i.* Time (or distance or branch length) is then measured by the expected number of amino acid substitutions per site. Given the rate matrix *Q,* the matrix of transition probabilities over time *t* can be calculated as $P(t) = \{p_{ij}(t)\} = e^{Qt}$, where $p_{ij}(t)$ is the probability that amino acid *i* changes into amino acid *j* after time *t.* A standard numerical algorithm can be used to calculate the eigenvalues and eigenvectors of *Q* to calculate *P*(*t*) (see Yang 1994*a*). The likelihood function can then be calculated using the approaches of Felsenstein (1981; see also Kishino, Miyata, and Hasegawa 1990; Goldman and Yang 1994) under models of one single rate for all sites and of Yang (1994*b*) under models of variable rates among sites. A numerical optimization algorithm is used to obtain maximum-likelihood estimates of parameters.

The simplest model of amino acid substitution is the Poisson model, which assumes an equal rate of change between any two amino acids (Bishop and Friday 1987). The equal-input (proportional) model assumes that the substitution rate is proportional to the frequency of the target amino acid; that is, $q_{ij} = \mu\pi_j$, where μ is a scale factor determined by equation (1). These models are known to fit real data poorly (see, e.g., Cao et al. 1994*a*, 1994*b*; Hashimoto and Hasegawa 1996) but are used here for comparison. Other models considered in this paper are of two kinds: empirical and mechanistic. The former summarizes the substitution pattern from a large amount of protein sequence data without considering what factors affect amino acid substitution. The latter attempts to describe the biological processes involved in amino acid substitution, such as mutational biases at the DNA level and selective constraints at the protein level.

Empirical Models of Dayhoff et al. and Jones et al.

The empirical model of amino acid substitution of Dayhoff, Schwartz, and Orcutt (1978) is implemented following Kishino, Miyata, and Hasegawa (1990). The updated matrix of Jones, Taylor, and Thornton (1992) is based on a much larger database. Counts of amino acid differences in pairwise sequence comparisons were provided by Dr. D. Jones and transformed into a matrix of substitution rates using the approach of Yang and Kumar (1996). The models are constructed to be time-reversible, such that the rate matrix is in the form $Q = S\Pi$, where $S = \{s_{ij}\}$ is a symmetrical matrix (with $s_{ij} = s_{ji}$ for any i and j), and $\Pi = \text{diag}\{\pi_1, \pi_2, \ldots, \pi_{20}\}$ (see, e.g., Yang 1994*a*; Zharkikh 1994). The elements of the rate matrix (s_{ij} and π_j) are fixed irrespective of the protein analyzed. When amino acid frequencies in the real data differ considerably from those predicted under the empirical models of Dayhoff, Schwartz, and Orcutt (1978) or Jones, Taylor, and Thornton (1992), use of amino acid frequencies as free parameters in the models provides a better fit. Then, only the s_{ij}'s are specified by the empirical models, and the π_j's are estimated from the data. These models, known as the Dayhoff-F and JTT-F models in the Molphy package (Adachi and Hasegawa 1996*a*), will be used in this paper.

The General Reversible Model

The general reversible Markov process model of amino acid substitution, referred to as ''REV,'' places the following constraint on the structure of the rate matrix:

$$\pi_i q_{ij} = \pi_j q_{ji}, \qquad \text{for any } i, j. \tag{2}$$

Equivalently, the rate can be written as $q_{ij} = s_{ij}\pi_j$, where $s_{ij} = s_{ji}$. The constraint reduces the number of free parameters in the rate matrix from 379 (=400 − 20 − 1) to 208 (=20 × 19/2 − 1 + 19). (Since only relative rates are considered, the number of free parameters is reduced by one.) The frequency parameters π_j are routinely estimated using the observed frequencies in the data, while elements in the symmetrical matrix, $S = \{s_{ij}\}$, are estimated by maximum likelihood. This parameter-rich model provides a basis with which other, simpler, models can be compared.

The general reversible model of amino acid substitution was first implemented by Adachi and Hasegawa (1996*a*, 1996*b*). The authors used the model to estimate the substitution pattern in mitochondrial proteins from a diverse range of species (including chicken, frog, fish, and lamprey as well as mammals) and supplied the estimated rate matrix as an empirical model in the `protml` program, i.e., the mtREV24 model (Adachi and Hasegawa 1996*a*). This empirical model is also used in our analysis. Comparison of this model with others will give us an indication as to whether it is appropriate to use a fixed-rate matrix such as mtREV24 in cases in which the data set is too small to estimate the rate matrix reliably or the estimation will be computationally too costly.

The Reversible Model Disallowing Changes at Multiple Codon Positions

This model, referred to as ''REV0,'' disallows instantaneous substitutions between amino acids that are different at two or three codon positions. The underlying assumption of the model is that mutations are independent at the three codon positions and that DNA level processes do not cause correlation among neighboring nucleotides. As a result, the probability that more than one codon position changes in a small time interval should be negligible. It should be noted that the probability of change from any amino acid to another is positive over any positive time interval; that is, $p_{ij}(t) > 0$ for any $t > 0$. The model merely assumes that amino acids differing at two or three codon positions cannot interchange without going through other, intermediate, amino acids. Under the universal genetic code, this model involves 93 (= 75 − 1 + 19) free parameters for the rate matrix, as there are 75 one-step amino acid pairs (pairs of amino acids separated by one codon position difference). Under the mammalian mitochondrial code, there are 70 one-step amino acid pairs, so the rate matrix involves 88 (= 70 − 1 + 19) free parameters.

Construction of an Amino Acid Substitution Model from a Markov Model of Codon Substitution

Suppose that substitutions between sense codons are described by a time-reversible Markov process with rate matrix $Q = \{q_{uv}\}$, where q_{uv} ($u \neq v$) is the rate of substitution from codon u to codon v. For the mammalian mitochondrial code, which has 60 sense codons, Q is a matrix of size 60 × 60. Since the process is reversible, we have $\pi_u q_{uv} = \pi_v q_{vu}$ for any u and v, and the rate can be written as $q_{uv} = s_{uv}\pi_v$, where $S = \{s_{uv}\}$ is a symmetrical matrix. Let $i = \text{aa}_u$ and $j = \text{aa}_v$ be the amino acids encoded by codons u and v, respectively. Here, we use the same symbols, q, s, and π for both codons and amino acids, and their subscripts indicate whether they refer to codons (u and v) or amino acids (i and j).

Let I and J represent the sets of codons that code for amino acids i and j, respectively; that is, $\text{aa}_u = i$ for $u \in I$, and $\text{aa}_v = j$ for $v \in J$. After combining synonymous codons that code for amino acid j into one state, we have $\pi_J = \sum_{v \in J} \pi_v$ to be the equilibrium frequency of amino acid j. Since substitutions to any synonymous codon (v) of amino acid j lead to amino acid j, the substitution rate from codon u to amino acid j is

$$q_{uJ} = \sum_{v \in J} q_{uv} = \sum_{v \in J} s_{uv}\pi_v. \tag{3}$$

The substitution rate from amino acid j to codon u is an average of substitution rates from synonymous codons (v) of amino acid j to codon u, weighted by the probability π_v/π_J that amino acid j is encoded by codon v:

$$q_{Ju} = \sum_{v \in J} \frac{\pi_v}{\pi_J} q_{vu} = \sum_{v \in J} \pi_u q_{uv}/\pi_J = \pi_u q_{uJ}/\pi_J. \tag{4}$$

The new Markov process, in which all synonymous codons for amino acid j are combined into one state, is time-reversible as $\pi_u q_{uJ} = \pi_J q_{Ju}$. Using similar arguments, we obtain the substitution rates between amino acids i and j by further combining synonymous codons for amino acid i into one state,

$$\begin{aligned} q_{JI} &= \sum_{u \in I} q_{Ju} = \sum_{u \in I} \pi_u q_{uJ}/\pi_J = \sum_{u \in I} \pi_u \sum_{v \in J} (s_{uv}\pi_v)/\pi_J \\ &= \sum_{u \in I} \sum_{v \in J} [\pi_u \pi_v s_{uv}/(\pi_I \pi_J)]\pi_I, \end{aligned} \tag{5}$$

$$q_{IJ} = \pi_J q_{JI}/\pi_I. \tag{6}$$

Let

$$s_{IJ} = \sum_{u \in I} \sum_{v \in J} \pi_u \pi_v s_{uv}/(\pi_I \pi_J), \tag{7}$$

and then $q_{IJ} = s_{IJ}\pi_J$ with $s_{IJ} = s_{JI}$. The resulting 20 × 20 matrix $Q = \{q_{IJ}\} = \{q_{ij}\}$ will be the rate matrix for a reversible Markov process of amino acid substitution. Since the codon frequencies are unavailable from the protein sequence data, we use equal frequencies for synonymous codons that code for the same amino acid to derive the symmetrical matrix S. Amino acid frequencies (π_i) are estimated using the observed frequencies in the data. Parameters in the substitution rate matrix include those in the codon substitution model and the amino acid frequencies.

Amino Acid Distance and Substitution Rate

A simplified version of the model of Goldman and Yang (1994) specifies the rate of substitution from codon u to codon v as

$$q_{uv} = \begin{cases} 0, & \text{if the two codons differ at more} \\ & \quad \text{than one position,} \\ \pi_v, & \text{for synonymous transversion,} \\ \kappa\pi_v, & \text{for synonymous transition,} \\ \omega\pi_v, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_v, & \text{for nonsynonymous transition,} \end{cases} \tag{8}$$

where κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio ($=d_N/d_S$ in the notation of Nei and Gojobori 1986), and π_v is the equilibrium frequency of codon v. When nucleotide (codon) sequences are analyzed, we use base frequencies at the three codon positions to calculate codon frequencies (π_v), with 9 ($=[3 \times (4 - 1)]$) free parameters used. When amino acid sequences are analyzed, π_v are calculated from amino acid frequencies under the assumption that synonymous codons are equally frequent, as mentioned above. Equation (8) is similar to the simulation model used by Gojobori (1983) or the model of Muse and Gaut (1994) and ignores differences in substitution rate between different amino acids. However, it is well known that different amino acids interchange at very different rates. In general, amino acids with similar physicochemical properties tend to replace each other more often than do dissimilar amino acids (e.g., Epstein 1964, 1967; Zuckerkandl and Pauling 1965; MacLachan 1972; Grantham 1974), presumably to maintain the conformation of the protein. Our purpose here is to compare several physicochemical distance measures between amino acids and determine what relationship between distance and rate best describes the real process of amino acid substitution. Thus, ω in equation (8) is replaced by ω_{ij}, where i ($=\text{aa}_u$) and j ($=\text{aa}_v$) are the two amino acids involved. The model specified by equation (8) is called the ‘‘equal-distance’’ model, as it is equivalent to using the same distance between any pair of amino acids. Following Miyata, Miyazawa, and Yasunaga (1979), we refer to ω_{ij} as the ‘‘acceptance’’ rate. When ω_{ij} varies with amino acids i and j, the overall d_N/d_S ratio can be estimated using the approach of Goldman and Yang (1994).

Several distance measures between amino acids have been suggested in the literature. Sneath (1966) used as many as 134 side chain properties but weighed them equally, so the distance does not appear useful. In this paper, we consider only the three properties used by Grantham (1974): composition (c), polarity (p) and volume (v). Composition is defined as the atomic weight ratio of noncarbon elements in end groups or rings to carbons in the side chain (Grantham 1974). Polarity and volume (size) are known to have a great impact on the folding of the protein, and are the only properties used by Epstein (1967) and Miyata, Miyazawa, and Yasunaga (1979). For each property x ($=c$, p, or v), we define the distance between amino acids i and j as $d_{ij} = |x_i - x_j|$. Distances based on composition range from 0 (for many amino acid pairs) to 2.75 (between Cys and any of Ala, Ile, Leu, Met, Phe, or Val), distances based on polarity range from 0 (for Arg-Gln) to 8.1 (for Leu-Asp), and distances based on volume range from 0.5 (for Pro-Ser) to 167 (for Gly-Tyr). Besides the three distances based on the three properties, we also use distance measures published by Grantham (1974) and Miyata, Miyazawa, and Yasunaga (1979). Grantham's distance is given as

$$\begin{aligned} d_{ij} = [&1.833(c_i - c_j)^2 + 0.1018(p_i - p_j)^2 \\ &+ 0.000399(v_i - v_j)^2]^{1/2}. \end{aligned} \tag{9}$$

and ranges from 5 for Leu-Ile to 215 for Cys-Trp. Miyata, Miyazawa, and Yasunaga's distance has a similar form,

$$d_{ij} = \sqrt{(p_i - p_j)^2/\sigma^2_{\Delta p} + (v_i - v_j)^2/\sigma^2_{\Delta v}}, \tag{10}$$

where $\sigma_{\Delta p}$ and $\sigma_{\Delta v}$ are the standard deviations of $|p_i - p_j|$ and $|v_i - v_j|$, respectively. The distance ranges from

0.06 for Pro-Ala to 5.13 for Gly-Trp. Grantham's and Miyata, Miyazawa, and Yasunaga's distance measures have a correlation coefficient of 0.76.

We consider a monotonic geometric relationship between distance (d_{ij}) and acceptance rate (ω_{ij})

$$\omega_{ij} = a \exp\{-bd_{ij}/d_{\max}\}, \qquad a \geq 0, \qquad (11)$$

as well as a linear relationship

$$\omega_{ij} = a(1 - bd_{ij}/d_{\max}), \qquad a \geq 0, b \leq 1. \qquad (12)$$

The maximum distance $d_{\max}$ is used to convert different distance measures to the same scale. The linear relationship is similar to a formula discussed by Miyata, Miyazawa, and Yasunaga (1979), while the formula used by Goldman and Yang (1994) is a special case of the geometric relationship using Grantham's distance, with $a = 1$ fixed. When $b = 0$, both relationships reduce to the equal-distance model of equation (8). When nucleotide (codon) sequences are analyzed, both parameters a and b can be estimated from the data. However, when amino acid sequences are analyzed, synonymous changes are invisible, and a becomes an inestimable scaling factor; in that case, parameter b alone will be estimated.

Different Types of Amino Acid Substitutions

Instead of using a mathematical function to describe the relationship between the acceptance rate (ω) and distance (d), amino acid interchanges can be classified into different types (groups) and assigned different rates. The acceptance rates can then be estimated from the data. For example, Epstein (1964) considered three types of amino acid interchanges: (I) those that do not alter polarity, (II) those between polar amino acids and glycine or alanine, and (III) those that alter polarity. Type I interchanges are expected to be frequent, while type III interchanges are expected to be rare. A few other classification schemes were discussed by Taylor (1986). In this paper, we use Grantham's distance to classify amino acid pairs into five groups, with the first group consisting of pairs of highly similar amino acids and the last group consisting of pairs of very different amino acids. The different groups of amino acid interchanges are assumed to have different ω's, which are estimated from the data.

The most general model of this kind is also implemented, which assigns an independent acceptance rate to each one-step amino acid pair; that is, each ω_{ij} is treated as a free parameter. This general model is useful for estimating the acceptance rates from the data without any constraint and provides a basis for comparison with other models.

Variable Substitution Rates Across Amino Acid Sites

The discrete-gamma model of Yang (1994*b*) was used to account for the variation of substitution rates across amino acid sites. Eight rate categories were used to approximate the (continuous-) gamma distribution, with each category having equal probability of occurrence (see Yang 1994*b* for implementation of the model).

Table 1
Log-Likelihood Values and Parameter Estimates Under Empirical Models for the Amino Acid Sequences of the Large Data Set

Model	p	ℓ	Tree Length	$\hat{\alpha}$	$\Delta\ell$
One rate for all sites					
Poisson	0	−42,774.51	1.534		
Equal input	19	−40,448.55	1.543		
Dayhoff-F	19	−37,380.84	1.603		
JTT-F	19	−36,922.84	1.600		
mtREV24-F	19	−36,372.15	1.605		
REV0	88	−36,844.94	1.786		
REV	208	−36,207.95	1.622		
Gamma rates for sites					
Poisson + G	1	−40,611.30	1.630	0.358	2,163.21
Equal input + G	20	−38,167.60	1.735	0.340	2,280.95
Dayhoff-F + G	20	−35,333.80	1.852	0.367	2,047.04
JTT-F + G	20	−34,910.37	1.819	0.372	2,012.47
mtREV24-F + G	20	−34,566.40	1.842	0.413	1,805.75
REV0 + G	89	−34,442.15	2.472	0.348	2,402.79
REV + G	209	−34,247.52	2.204	0.372	1,960.43

NOTE.—p is the number of parameters in the substitution model not including the 37 branch lengths. α is the gamma shape parameter, and $\Delta\ell$ is the log-likelihood difference between the single-rate and gamma-rates models. Tree length is the expected number of amino acid substitutions per site along the tree.

Results

Analysis of the Large Data Set

Rate Variation Among Sites

The empirical models were used in maximum-likelihood analyses of the protein sequences in the large data set of 20 species. Log-likelihood values and estimates of the tree length (sum of branch lengths along the tree) are listed in table 1. For the Dayhoff-F and JTT-F models, amino acid frequencies were treated as free parameters and estimated by the observed frequencies in the sequence data. The substitution models were used assuming either a single rate for all sites or gamma-distributed rates among sites (Yang 1994*b*). Estimates of the gamma shape parameter (α) are shown, as are the log-likelihood differences ($\Delta\ell$) between the single-rate and gamma-rates models (table 1). The single-rate model is a special case of the gamma model with $\alpha = \infty$ (see, e.g., Yang 1994*b*). Thus, $2\Delta\ell$ can be compared with the χ^2 distribution with one degree of freedom to test for constancy of rates among sites. The observed statistics ($2\Delta\ell$) range from 3,611 to 4,805 among the substitution models of table 1 and are all much greater than the critical value $\chi^2_{1\%} = 6.64$. The gamma model thus provides a significantly better fit to the mitochondrial protein sequences than does the single-rate model. There is no doubt that substitution rates are highly variable among amino acid sites (Uzzell and Corbin 1971; Golding 1983; Reeves 1992). The estimated α for the gamma model ranges from 0.34 to 0.37, indicating severe rate variation among sites.

The tree length, that is, the total number of amino acid substitutions per site along the tree, ranges from 1.5 to 1.8 under models of a single rate for all sites, and from 1.6 to 2.5 under the gamma models. In general, simple models tend to underestimate sequence distances or tree lengths. Furthermore, the rate variation among sites has a greater effect on estimation of branch lengths and tree lengths than the substitution pattern. Those results are consistent with early observations based on nucleotide sequences. For example, it was noted that ignoring either the rate variation among sites or the transition-transversion bias leads to underestimation of sequence distances or branch lengths in a phylogenetic tree, but the rate variation has a much more serious effect than the transition bias (e.g., Yang, Goldman, and Friday 1994).

It is noted that the relative performances of the six different amino acid substitution models did not change regardless of whether the gamma model was used to account for variable rates among sites (table 1). Thus, some of the analyses in this paper comparing different models (patterns) of amino acid substitution are performed assuming a single rate for all sites. In such cases, it is expected that accounting for the rate variation among sites would not change our conclusions concerning the substitution pattern, although it would lead to significant improvement of the fit of the models to data.

Empirical Models of Amino Acid Substitution

Consistent with previous analyses of mitochondrial and nuclear proteins using empirical models (e.g., Cao et al. 1994*a*, 1994*b*), the log-likelihood values indicating the goodness of fit of the models are in the following order: Poisson < equal-input < Dayhoff-F or JTT-F (table 1). The empirical models of Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992) fit the mitochondrial data much better than the equal-input model. The log-likelihood differences ($\Delta\ell$) range from 2,834 to 3,526 in such comparisons. The equal-input model is unrealistic, as it ignores differences in substitution rate between amino acids. JTT-F has higher log-likelihood values than Dayhoff-F. This appears to be the case for most data sets (see, e.g., Cao et al. 1994*a*, 1994*b*), although exceptions do exist (unpublished data).

The general reversible model (REV) is significantly better than Dayhoff-F or JTT-F. For example, twice the log-likelihood difference between REV and JTT-F is $2\Delta\ell = 1{,}429.78$ under the single-rate model and $2\Delta\ell = 1{,}325.70$ under the gamma-rates model. The critical value for those comparisons is 237.15 with df = 189, and JTT-F is rejected irrespective of the assumption about rates at sites. Relative substitution rates under different models are plotted in figure 2. The three amino acid pairs with the highest rates differ among the models and are Asp-Glu, Asn-Asp, and Ile-Val under Dayhoff-F; Ile-Val, Asp-Glu, and Lys-Arg under JTT-F; and Ile-Val, Tyr-His, and Asn-Asp under REV or REV + G. Two important factors appear to account for the poor fit of the Dayhoff-F and JTT-F models to those mitochondrial data. The first is the use of different genetic codes in the nuclear and mitochondrial genomes, as pointed out by Adachi and Hasegawa (1996*b*). For example, the Dayhoff-F and JTT-F models were constructed from nuclear proteins and predict frequent interchanges between arginine and lysine (fig. 2), which are separated by one substitution at the second codon position in the nuclear code. In the mitochondrial code, the two amino acids are separated by at least two codon position differences, and substitutions between them are rare (see rate estimates under REV and REV + G in fig. 2). The second factor appears to be the inadequacies of the estimation procedures used to construct the empirical matrices of Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992). For example, rates under Dayhoff-F and JTT-F are much more homogeneous than those under REV, which are in turn more homogeneous than those under REV + G (fig. 2). In previous analyses of nucleotide sequences, it was noted that parsimony (which was used to construct the Dayhoff-F matrix), as well as likelihood assuming one rate for all sites, was ineffective in correcting for multiple hits and underestimated the transition-transversion bias (Wakeley 1994; Yang, Goldman, and Friday 1994). A similar interpretation should apply to the rate differences among models observed here. Of course, rates estimated under the REV model may involve biases and sampling errors too. However, it may be expected that only small rate estimates (for infrequent amino acid interchanges) involve substantial errors and that rates for frequent interchanges are estimated reliably. It appears feasible to use a parameter-rich model such as REV to estimate the amino acid substitution pattern when a large amount of protein sequence data are available.

We also note that the mtREV24 model of Adachi and Hasegawa (1996*a*, 1996*b*) gave much better fit to the mitochondrial proteins than did the models of Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992). While this result is not surprising, as the mtREV24 rate matrix was originally derived from mitochondrial proteins, it does provide justification for the use of the model in analyzing mitochondrial data, even if the species in the data set are not the same as those used by Adachi and Hasegawa (1996*a*).

Comparison between REV0 and REV constitutes a test of the hypothesis that amino acid (codon) substitutions proceed in a stepwise manner, with each step involving a change at only one codon position. The test statistic is $2\Delta\ell = 2 \times [-36{,}207.95 - (-36{,}844.94)] = 1{,}273.98$ under the single-rate model and $2\Delta\ell = 2 \times [-34{,}247.52 - (-34{,}442.15)] = 389.26$ under the gamma model (table 1). The critical value is $\chi^2_{1\%} = 158.95$ with df = 120. Accounting for the among-site rate variation has improved the fit of the REV0 model considerably such that the likelihood difference between REV0 and REV under the gamma model is much smaller than that under the single-rate model. However, REV provides a significantly better fit than REV0 irrespective of the assumption about substitution rates among sites. The result suggests that certain processes at the DNA level may have caused interdependence of substitutions at the three codon positions. Note that correlation among codon positions caused by the genetic code and by selective constraints at the protein level is accounted for by both REV0 and REV and does not contribute to the significant difference between the two models. Possible factors are mutations affecting more than one nucleotide site, compensatory nucleotide substitutions, and selec-

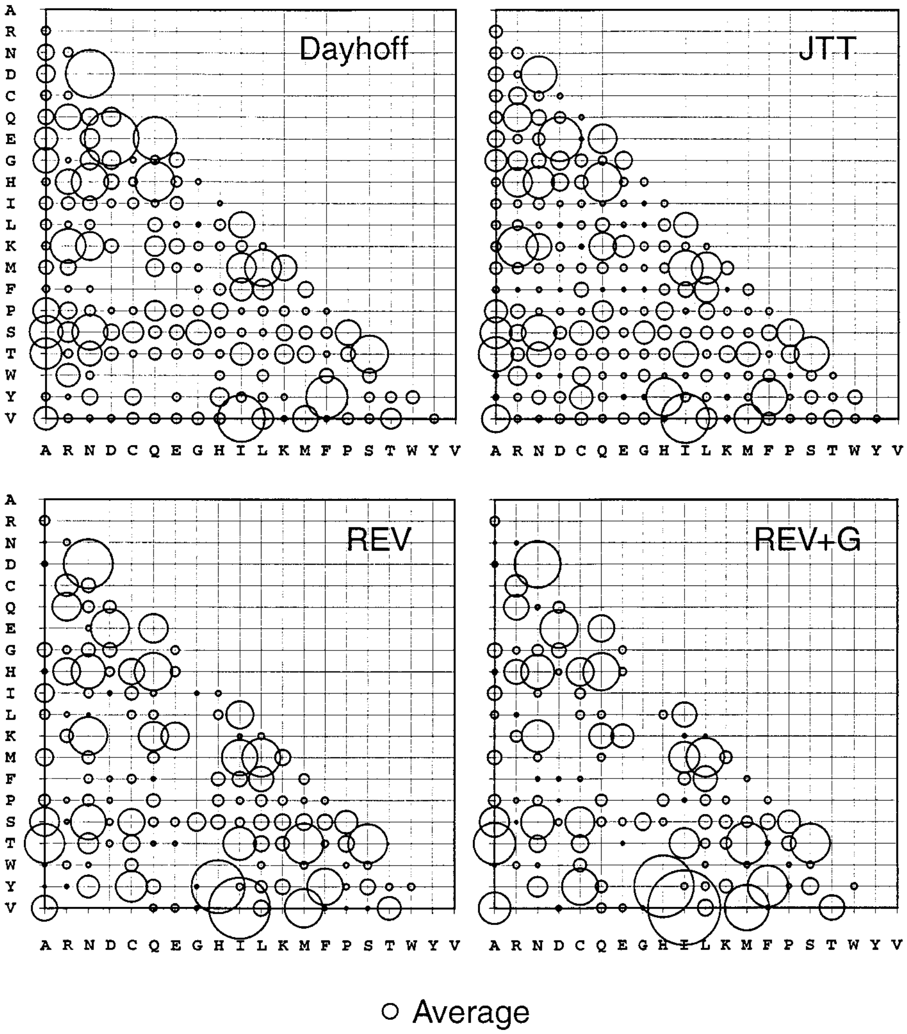

FIG. 2.—Plots of relative substitution rates (s_{ij}) between amino acids under different models. The sizes (areas) of bubbles are proportional to the substitution rates and are comparable in different plots. For example, the highest substitution rate (between Ile and Val) under REV + G is 22 times as high as the average rate, while rates between many amino acid pairs are close to zero. Rates under Dayhoff-F and JTT-F are fixed, while those under REV and REV + G are estimated by maximum likelihood from the protein sequences of the large data set. The Poisson and equal-input models predict the same rate (s_{ij}), the average shown at the bottom of the figure, for any pair of amino acids.

tive pressures at the DNA level (Krzywicki and Slonimski 1968; Schöniger and von Haeseler 1994). It will be interesting to analyze more data, especially gene sequences from the nuclear and chloroplast genomes, to find out whether the pattern holds for other genomes and genes and what factors are responsible.

Analysis of the Small Data Set

Empirical Models of Amino Acid Substitution

More detailed analyses were performed on the small data set using both amino acid and nucleotide sequences. The relative performance of the empirical models applied to the amino acid sequences (table 2) shows the same pattern as for the large data set. The log-likelihood differences among models are much smaller, apparently due to the reduced discriminating power of the small data set. However, conclusions reached in analysis of the large data set are all significantly supported in the small data set. For instance, the log-likelihood differences between the one-rate and corresponding gamma-rates models range from $\Delta\ell = 134.22$ under JTT-F to $\Delta\ell = 171.70$ under the equal-input model (table 2). Comparison of twice the log-likelihood differences ($2\Delta\ell$) with $\chi^2_{1\%} = 6.63$ (df = 1) suggests that rate constancy among sites is rejected no matter which substitution model is used. Furthermore, the log-likelihood difference between REV and JTT-F is $\Delta\ell = 171.47$ when rates are assumed to be identical among sites, and it is 185.49 when they are assumed to be gamma distributed. Note that the 189 rate parameters in the REV models were not estimated from the small data set, and estimates obtained from the large data set were used as fixed constants. The likelihood values for the REV models of table 2 are thus underestimated. Even so, JTT-F is rejected in favor of REV ($2\Delta\ell$ compared with $\chi^2_{1\%} = 237.15$ [df = 189]). The relative performance of the mtREV24 model (Adachi and Hasegawa 1996*a*) is also similar to that for the large data set.

Table 2
Analyses of Amino Acid Sequences of the Small Data Set: Empirical Models

Model	p	ℓ	Tree Length	$\hat{\alpha}$
One rate for all sites				
Poisson	0	−16,566.60	0.275	
Equal input	19	−15,503.26	0.275	
Dayhoff-F	19	−14,904.33	0.279	
JTT-F	19	−14,717.98	0.279	
mtREV24-F	19	−14,687.36	0.280	
REV (rate matrix Q from REV + G of table 1)	208	−14,546.51	0.285	
Gamma rates for sites				
Poisson + G	1	−16,399.92	0.291	0.287
Equal input + G	20	−15,331.56	0.295	0.278
Dayhoff-F + G	20	−14,765.25	0.294	0.354
JTT-F + G	20	−14,58376	0.291	0.368
mtREV24-F + G	20	−14,558.67	0.298	0.392
REV + G (rate matrix Q from REV + G of table 1)	209	−14,398.27	0.318	0.368

NOTE.—p is the number of parameters in the substitution model not including the 11 branch lengths. Tree length is the expected number of amino acid substitutions per site along the tree. α is the gamma shape parameter.

Amino Acid Distance and Substitution Rate

Amino acid sequences of the small data set were also analyzed using the codon-based mechanistic models, with one single rate for all sites assumed (table 3). The equal-distance model (eq. 8) fits the data much better than the equal-input model of table 2. While the two models are not nested and a χ^2 approximation cannot be used to compare them, the log-likelihood difference between the two models ($\Delta\ell = 785.04$) is huge. Both models account for different amino acid frequencies. However, the equal-distance model accounts for the ''mutational distance'' between amino acids determined by the genetic code and the transition-transversion bias, while the equal-input model ignores the mutational distance. Indeed, the equal-distance model is much better than Dayhoff-F and is as good as JTT-F, indicating that the genetic code and mutational biases have a great impact on amino acid substitution rates.

Five physicochemical distances between amino acids are considered in table 3: Grantham's, Miyata, Miyazawa, and Yasunaga's, and the three distances constructed using the side chain composition, polarity, and volume of amino acids, respectively. Both geometric and linear relationships were used for each distance measure, and they gave similar performances (table 3). Use of any of the five distance measures leads to a better fit to data than does the equal-distance model (table 3), suggesting that those properties do influence amino acid substitution. All mechanistic models considered in table 3 gave better fits to data than the empirical models of Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992) (table 2). The distance based on composition gave the poorest fit, and we note that this property was not used by Miyata, Miyazawa, and Yasunaga (1979). Many amino acids are not distinguished by this property; for example, Ala, Ile, Leu, Met, Phe,

Table 3
Analyses of Amino Acid Sequences of the Small Data Set: Codon-Based Mechanistic Models

Model	p	ℓ	$\hat{\kappa}$	$\hat{b}$
Equal distance	20	−14,718.22	9.157	0
Geometric (eq. 11)				
Composition (c)	21	−14,713.78	9.357	0.607
Polarity (p)	21	−14,676.32	10.256	1.623
Volume (v)	21	−14,668.27	11.435	2.291
Grantham's distance (cpv)	21	−14,681.03	10.555	1.629
Miyata, Miyazawa, and Yasunaga's distance (pv)	21	−14,658.32	11.050	2.093
Linear (eq. 12)				
Composition (c)	21	−14,713.57	9.341	0.535
Polarity (p)	21	−14,673.54	9.817	1.000
Volume (v)	21	−14,671.12	10.279	1.000
Grantham's distance (cpv)	21	−14,679.20	10.077	0.954
Miyata, Miyazawa, and Yasunaga's distance (pv)	21	−14,656.58	10.173	1.000

NOTE.—p is the number of parameters in the substitution model not including the 11 branch lengths. Parameter b is defined in equations (11) and (12), while a in those equations is inestimable from amino acid sequences.

Table 4
Analyses of the Nucleotide (Codon) Sequences of the Small Data Set

Model	p	ℓ	$\hat{\kappa}$	$\hat{a}$	$\hat{b}$	d_N/d_S
Equal distance	11	−29,967.86	14.249	0.041	0	*a*
Geometric (eq. 11)						
Composition (c)	12	−29,961.86	14.299	0.046	0.699	0.041
Polarity (p)	12	−29,917.48	14.540	0.062	1.838	0.042
Volume (v)	12	−29,900.76	15.173	0.072	2.557	0.041
Grantham's distance (cpv)	12	−29,911.54	14.773	0.076	2.065	0.042
Miyata, Miyazawa, and Yasunaga's distance (pv)	12	−29,890.18	14.901	0.080	2.399	0.042
Linear (eq. 12)						
Composition (c)	12	−29,961.34	14.294	0.046	0.617	0.041
Polarity (p)	12	−29,917.34	14.353	0.055	1.000	0.042
Volume (v)	12	−29,909.83	14.672	0.055	1.000	0.041
Grantham's distance (cpv)	12	−29,915.68	14.518	0.061	0.998	0.042
Miyata, Miyazawa, and Yasunaga's distance (pv)	12	−29,895.91	14.544	0.060	1.000	0.042

NOTE.—p is the number of parameters in the model not including the 11 branch lengths in the tree. Parameters a and b are defined in equations (11) and (12), and κ is the transition/transversion rate ratio. In each model, base frequencies at the three codon positions are used to calculate codon frequencies, with $3 \times (4 - 1) = 9$ free parameters used.

and Val were all assigned the value zero. The ranking of the distance measures based on the three properties is $c < p < v$, and volume (size) appears to have the greatest influence on amino acid substitutions in mitochondrial proteins. Grantham's combined distance based on all three properties has a performance between distances based on c and p (table 3). Miyata, Miyazawa, and Yasunaga's distance gave the best performance among all five measures. This is probably because Miyata, Miyazawa, and Yasunaga (1979) constructed their distances to achieve a high correlation with substitution rates between one-step amino acid pairs, while Grantham (1974) considered all possible pairs. As pointed out by Miyata, Miyazawa, and Yasunaga (1979), amino acids separated by two or three codon position differences are unlikely to interchange even if they are chemically similar.

The codon substitution models were also applied to the codon (nucleotide) sequences of the seven primate species (table 4). The relative performances of different models are largely the same as those found when the models were applied to the amino acid sequences (table 3). The ranking of the distances based on the three properties is again $c < p < v$, and Grantham's distance lies between p and v, slightly better than when it was applied to the amino acid sequences. Miyata, Miyazawa, and Yasunaga's distance gave the best performance among the five distance measures examined. The geometric and linear relationships performed similarly. The best combination of distance and distance–rate relationship is Miyata, Miyazawa, and Yasunaga's distance with the geometric relationship (table 4). Use of nucleotide sequences makes it possible to estimate model parameters reliably. In particular, estimates of the overall d_N/d_S ratio obtained using the approach of Goldman and Yang (1994) are between 0.041 and 0.042, almost identical among models (table 4). Estimates of the transition/transversion rate ratio (κ) are very similar among models and range from 14 to 15, indicating the strong transition bias in the mitochondrial genome. These estimates suggest that the transition/transversion rate ratio was slightly underestimated from protein sequence data (table 3). Estimates of parameter b from the amino acid (table 3) and nucleotide (table 4) sequences are similar. The results suggest that amino acid sequences can be safely used to compare different codon substitution models, although nucleotide (codon) sequences are needed to estimate the d_N/d_S rate ratio.

It may be noted that the model of Goldman and Yang (1994) is equivalent to the geometric relationship with Grantham's distance with parameter $a = 1$ fixed. The log-likelihood value under this model is −30,774.01, much lower than that under the equal-distance model (table 4). The estimate of a without fixing it at one is 0.076 (table 4), and the assumption of $a = 1$ is clearly unrealistic.

Different Types of Amino Acid Substitution

We group amino acid pairs (interchanges) into five classes based on Grantham's distance measure (table 5). The first group includes nine highly similar pairs, with distances <26, the second group includes more different pairs, with distances from 26 to 47, while the fifth group includes very different pairs, with distances >89. The five groups are assumed to have different acceptance rates (ω), which are estimated from the nucleotide sequences of the small data set (table 5). The log-likelihood value under this model is −29,879.95, with $\hat{\kappa} = 14.446$ and $d_N/d_S = 0.042$. The model fits the data better than the best model of table 4, i.e., the geometric relationship with Miyata, Miyazawa, and Yasunaga's distance. The two models are not nested, and a χ^2 approximation to the likelihood ratio statistic cannot be applied. Nevertheless, we note that the log-likelihood difference for this comparison ($\Delta\ell = 10.23$) is not great.

We also applied the general model that assumes an independent acceptance rate for each one-step amino acid pair to the nucleotide sequences of the small data

Table 5
Estimates of Acceptance Rates for Different Types of Amino Acid Interchanges from Nucleotide Sequences of the Small Data Set

Class (distance range)	Amino Acid Pairs	ω
1 (5-26)	DN, HQ, LI, MI, ML, FI, FL, YF, VM	0.045
2 (26–47)	QR, ED, EQ, HR, PA, SN, TP, VI, VL	0.066
3 (47–68)	GA, KQ, KE, SG, TA, TN, TS, WL, VA, VF	0.065
4 (68–89)	HN, HD, PQ, PH, SP, TK, TM, YH	0.040
5 (89–215)	All other pairs with one position difference	0.020

NOTE.—Amino acid interchanges are grouped into five classes according to Grantham's distances. Amino acids are represented by the one-letter codes.

set. As there are 70 such pairs by the mitochondrial code, this model involves 80 free parameters (70 acceptance rates, one transition/transversion rate ratio κ, and nine parameters for base frequencies at the three codon positions). The log-likelihood value is −29,667.57, with $\hat{\kappa} = 18.356$ and the overall $d_N/d_S = 0.038$. The five highest acceptance rates are 0.321 for Thr-Ser, 0.196 for Ser-Ala, 0.188 for Met-Leu, 0.178 for Asp-Glu, and 0.147 for Ser-Cys. Since all models in tables 4 and 5 are special cases of this general model, we use the likelihood ratio test to examine the fit of the simpler models. Such comparisons suggest that all models in tables 4 and 5 should be rejected. For example, the general model assumes 70 different ω parameters for the 70 one-step amino acid pairs, while the equal-distance model assumes 1 ω parameter. Twice the log-likelihood difference between the two models is $2\Delta\ell = 2 \times [-29{,}667.57 - (-29{,}967.86)] = 600.59$, much greater than the critical value $\chi^2_{1\%} = 99.23$ with df = 70 − 1 = 69. The equal-distance model is thus rejected, and, indeed, the acceptance rates for different amino acid pairs are different. Furthermore, the best model in table 4, that is, Miyata, Miyazawa, and Yasunaga's distance with the geometric relationship, is also rejected when compared with the general model. Twice the log-likelihood difference between the two models is $2\Delta\ell = 2 \times [-29{,}667.57 - (-29{,}890.18)] = 445.21$. The critical value is $\chi^2_{1\%} = 98.02$ with df = 80 − 12 = 68. Figure 3 plots the acceptance rates (ω_{ij}) estimated under the general model against Miyata, Miyazawa, and Yasunaga's distances. While the acceptance rates between similar amino acids tend to be greater than those between dissimilar amino acids, the relationship between rate and distance does not appear to be describable by a simple mathematical function.

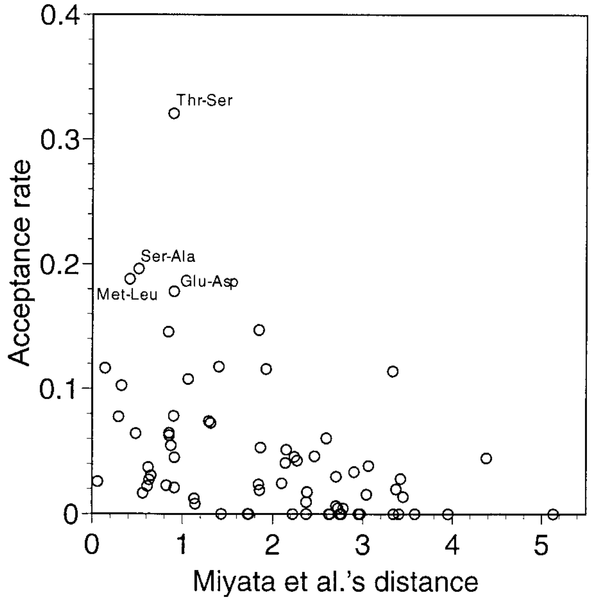

FIG. 3.—Amino acid "acceptance" rate (ω_{ij}) plotted against Miyata, Miyazawa, and Yasunaga's distance (d_{ij}). The rates were estimated by maximum likelihood from the nucleotide (codon) sequences of the small data set. Amino acid pairs separated by more than one codon position difference are not used in the plot, as they are assumed to have zero instantaneous rates of change.

Discussion

Analyses of numerous DNA sequences suggest that transition-transversion bias and nucleotide frequency biases are two prominent features of DNA sequence evolution. The gain upon adding extra complexities in the nucleotide substitution pattern is often noted to be minor (e.g., Yang 1994*a*). Since both these features of nucleotide substitution are considered in the codon substitution models examined in this paper, the models should most likely be improved by a better specification of the acceptance rate (ω_{ij}; see eq. 8). Physicochemical properties of amino acids, especially polarity and size, are known to affect the conformation of the protein and, thus, the substitution rate between amino acids. In this paper, we compared two empirical distance–rate relationships in combination with five physicochemical distance measures. The results suggest that codon substitution models using simple amino acid distance measures to specify acceptance rates fit the mitochondrial protein sequence data better than empirical models of Dayhoff, Schwartz, and Orcutt (1978) and Jones, Taylor, and Thornton (1992), constructed from large databases. Nevertheless, the codon-based models are rejected when compared with the general model, which treats all acceptance rates as free parameters. In particular, the relationship between distance and rate does not appear to permit a simple mathematical description. More work is needed to better understand the mechanisms of protein sequence evolution, and reliable estimates of substitution (s_{ij}) or acceptance (ω_{ij}) rates, which can be obtained using methods developed in this paper, may be very useful in this endeavor.

Gillespie (1991, pp. 40–44) pointed out that the neutral and selectionist theories of molecular evolution make different predictions about the relationship between substitution rate and physicochemical distance. Instead of the strictly decreasing relationships considered here (eqs. 11 and 12), which are consistent with the neutral prediction (Kimura 1983, p. 159), the rate may have a peak at an intermediate distance if both negative and positive selection operate on the protein. A plot (not shown) of our maximum-likelihood estimates of substitution rates (s_{ij}'s) against Grantham's distance shows a pattern similar to that shown in figure 1.12 of Gillespie (1991). (The acceptance rates of fig. 3 appear better suited for this purpose.) As pointed out by Gillespie, the highest rates do not occur between the most similar amino acids but, rather, occur at intermediate distances (see fig. 3). However, there does not appear to be a strong relationship between rate and distance. To test Gillespie's (1991) hypothesis of natural selection more rigorously, we extend the model of equation (11) so that the highest substitution rate occurs at a nonzero ''optimum'' distance c, and we apply the model to nucleotide sequences of the small data set. The acceptance rate is specified as

$$\omega_{ij} = a \exp\{-b|d_{ij}/d_{\max} - c|\}, \qquad (13)$$

where a, b, and c are parameters estimated from the data. When $c = 0$, the model reduces to the strictly decreasing relationship of equation (11). The log-likelihood value under this model with d_{ij} given by Grantham's distance is $\ell = -29{,}911.54$, with parameter estimates $\hat{\kappa} = 14.773$, $\hat{a} = 0.072$, $\hat{b} = 2.065$, and $\hat{c} = 0.022$. This model does not fit the data any better than the monotonic relationship (table 4). The same model using Miyata, Miyazawa, and Yasunaga's distance gave $\ell = -29{,}889.94$, $\hat{\kappa} = 14.888$, $\hat{a} = 0.075$, $\hat{b} = 2.408$, and $\hat{c} = 0.027$. The improvement in the model's fit upon adding the extra parameter c is insignificant (see table 4). We also considered a model similar to equation (13) but with a linear relationship being used or with parameter $b = 1$ fixed. The results are somewhat similar and are not presented here. In sum, these likelihood ratio tests do not provide support for Gillespie's (1991) argument.

Acknowledgments

We thank David Pollock and two anonymous referees for many useful comments. This study is supported by grants from the Ministry of Education, Science, Sports and Culture of Japan to Z.Y. and M.H., NIH grant GM40282 to M. Slatkin, and a fellowship from the Danish Research Council to R.N.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1992. MOLPHY: programs for molecular phylogenetics, I.—PROTML: maximum likelihood inference of protein phylogeny. Comput. Sci. Monogr. **27**:1–77.

———. 1996*a*. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. **28**:1–150.

———. 1996*b*. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

BISHOP, M. J., and A. E. FRIDAY. 1985. Evolutionary trees from nucleic acid and protein sequences. Proc. R. Soc. Lond. B. Biol. Sci. **226**:271–302.

———. 1987. Tetropad relationships: the molecular evidence. Pp. 123–139 *in* C. PATTERSON, ed. Molecules and morphology in evolution: conflict or compromise? Cambridge University Press, Cambridge, England.

BROWN, J. R., and W. F. DOOLITTLE. 1995. Root of the universal tree of life based on ancient aminoacyl–tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. USA **92**: 2441–2445.

CAO, Y., J. ADACHI, A. JANKE, S. PAABO, and M. HASEGAWA. 1994*a*. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. **39**:519–527.

CAO, Y., J. ADACHI, T. YANO, and M. HASEGAWA. 1994*b*. Phylogenetic place of guinea pigs: no support of the rodent polyphyly hypothesis from maximum likelihood analyses of multiple protein sequences. Mol. Biol. Evol. **11**:593–604.

CAO, Y., A. JANKE, P. J. WADDELL, M. WESTERMAN, O. TAKENAKA, S. MURATA, N. OKADA, S. PAABO, and M. HASEGAWA. 1998. Conflict amongst individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. (in press).

COATES, M., and S. STONE. 1981. Simulation of protein evolution by random fixation of allowed codons. J. Mol. Evol. **17**:311–328.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* Atlas of protein sequence and structure. Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

EPSTEIN, C. J. 1964. Relation of protein evolution to tertiary structure. Nature **203**:1350–1352.

———. 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature **215**:355–359.

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

GILLESPIE, J. H. 1991. The courses of molecular evolution. Oxford University Press, Oxford.

GOJOBORI, T. 1983. Codon substitution in evolution and the ''saturation'' of synonymous changes. Genetics **105**:1011–1027.

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

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. **185**:862–864.

HASEGAWA, M., and T. HASHIMOTO. 1993. Ribosomal RNA trees misleading? Nature **361**:23.

HASHIMOTO, T., and M. HASEGAWA. 1996. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. Adv. Biophys. **32**:73–120.

IWABE, N., K. KUMA, M. HASEGAWA, S. OSAWA, and T. MIYATA. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**:9355–9359.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

Jorré, R. P., and N. Curnow. 1975. A model for the evolution of the proteins. Biochemie **57**:1141–1146.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

Krzywicki, A., and P. P. Slonimski. 1968. Formal analysis of protein sequences. II. Method for structural studies of homologous proteins and amino acid substitutions in cytochrome *c*. J. Theor. Biol. **21**:305–330.

MacLachlan, A. D. 1972. Repeating sequences and gene duplication in proteins. J. Mol. Evol. **64**:417–437.

Miyata, T., S. Miyazawa, and T. Yasunaga. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. **12**:219–236.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome. Mol. Biol. Evol. **11**:715–724.

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

Reeves, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. **35**:17–31.

Schöniger, M., and A. von Haeseler. 1994. A stochastic model and the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3**:240–247.

Sneath, P. H. A. 1966. Relations between chemical structures and biological activity in peptides. J. Theor. Biol. **12**:157–195.

Taylor, W. R. 1986. The classification of amino acid conservation. J. Theor. Biol. **119**:205–218.

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

Uzzell, T., and K. W. Corbin. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**: 1089–1096.

Wakeley, J. 1994. Substitution rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:436–442.

Wilbur, W. J. 1985. On the PAM matrix model of protein evolution. Mol. Biol. Evol. **2**:434–447.

Yang, Z. 1994*a*. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

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

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.

Yang, Z., and S. Kumar. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13**: 650–659.

Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. **12**:451–458.

Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. **39**:315–329.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–116 *in* V. Bryson and H. J. Vogel, eds. Evolving genes and proteins. Academic Press, New York.

Axel Meyer, reviewing editor

Accepted August 17, 1998