J Mol Evol (1999) 48:86–93

JOURNAL OF MOLECULAR EVOLUTION

Variation in the Pattern of Nucleotide Substitution Across Sites

John P. Huelsenbeck,[1] Rasmus Nielsen[2,*]

[1] Laboratory of Molecular Systematics, Smithsonian Museum Support Center, 4210 Silver Hill Road, Suitland, MD 20746, USA
[2] Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

Received: 4 November 1997 / Accepted: 19 May 1998

Abstract. A model of nucleotide substitution that allows the transition/transversion rate bias to vary across sites was constructed. We examined the fit of this model using likelihood-ratio tests by analyzing 13 protein coding genes and 1 pseudogene. Likelihood-ratio testing indicated that a model that allows variation in the transition/transversion rate bias across sites provided a significant improvement in fit for most protein coding genes but not for the pseudogene. When the analysis was repeated with parameters estimated separately for first, second, and third codon positions, strong heterogeneity was uncovered for the first and second codon positions; the variation in the transition/transversion rate was generally weaker at the third codon position. The transition rate bias and branch lengths are underestimated when variation in the transition/transversion rate was not accommodated, suggesting that it may be important to accommodate variation in the pattern of nucleotide substitution for accurate estimation of evolutionary parameters.

Key words: Likelihood-ratio testing — Maximum likelihood — Molecular evolution — Molecular systematics — Phylogeny

Introduction

Most phylogenetic methods assume that the pattern of nucleotide substitution remains constant across the sites of a gene or across subsets of a gene. For example, parsimony assumes that the weight matrix assigning costs to different nucleotide changes remains constant and maximum-likelihood and distance methods assume that the matrix specifying the instantaneous rate of change from one nucleotide to another remains constant. Several authors have relaxed the constraint that the rates at different sites are constant by differentially weighting site positions (e.g., Farris 1969; Goloboff 1993) or by assuming that the rate at a site is a random variable drawn from a distribution [e.g., for distance and maximum-likelihood methods that incorporate among-site rate variation (Felsenstein 1981; Golding 1983; Jin and Nei 1990; Yang 1993)]. However, allowing rate variation across sites only changes the branch lengths of the tree; it does not relax the constraint that the same pattern of substitution (or relative rates of substitution among different nucleotides) applies across sites.

Only a few studies have examined how nucleotide substitutions vary across sites. That the rates of substitution vary across sites has been well established [see Yang (1996a) for a review of rate variation across the sites of a sequence]. Rate variation across sites is thought to reflect variation in the selection acting on different nucleotide positions because of functional and structural requirements for the protein; rate variation does appear to be smaller in genes that are not affected strongly by selection [e.g., adding among-site rate variation to the substitution model for a pseudogene has less effect on the fit of the substitution model to the data than adding among-site rate variation for coding genes (Yang et al. 1994)]. Yang (1996b) examined across-site heterogeneity in other substitution parameters. He estimated substitution parameters such as branch lengths, base frequen-

* *Present address:* Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA
Correspondence to: J.P. Huelsenbeck; *e-mail:* johnh@onyx.si.edu

cies, and transition/transversion rate ratio separately for first, second, and third codon positions of mitochondrial protein coding genes and found that all parameters varied significantly among codon positions. However, Yang (1996b) did not investigate whether the pattern of nucleotide substitution varied within codon positions.

In this study, we test the assumption that the pattern of nucleotide substitution is constant across sites. We do this by developing a general model of DNA substitution that allows the transition/transversion rate ratio to vary across sites. We examine the fit of this model using likelihood-ratio tests on 13 protein coding mitochondrial genes and one pseudogene.

Methods

Model of DNA Substitution. In our analysis, we assume that DNA sequences from homologous regions are available. Let $\mathbf{X} = \{x_{ij}\}$ be the aligned nucleotide sequences, where $i = 1, 2, \ldots, s$ and $j = 1, 2, \ldots, c$; s is the number of sequences sampled; and c is the number of nucleotide sites per sequence. Each column in the data matrix, $\mathbf{x}_j = \{x_{1j}, \ldots, x_{sj}\}$, specifies the nucleotides for the s sequences at the jth site.

We assume that DNA substitutions follow a time-homogeneous Markov process. As a starting point, we use the model of DNA substitution proposed by Hasegawa et al. (1985; see also Hasegawa et al. 1984) and designated the HKY85 model. The HKY85 model allows different base frequencies and a transition/transversion substitution bias. The instantaneous rate of substitution for the HKY85 model is specified by the rate matrix

$$\mathbf{Q} = \{q_{ij}\} = \mu \begin{pmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{pmatrix} \quad (1)$$

where q_{ij} ($i \neq j$) is the rate of substitution from nucleotide i to nucleotide j and the rows and columns are in the order A, C, G, T. The diagonals of **Q** are specified by the mathematical requirement that each row sums to 0. The factor μ is the mean instantaneous rate of substitution. This scaling factor is chosen so that the average rate of substitution is 1. Hence, the branch lengths of the phylogenetic tree are measured in terms of expected number of substitutions per site (v). The equilibrium distribution of nucleotides is given by $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$. The transition/transversion rate ratio is specified by the parameter κ; $\kappa > 1$ indicates that transitions occur at a higher rate than transversions. The free parameters of the HKY85 model (κ and π) are contained in a vector $\Theta_1 = \{\kappa, \pi\}$. The transition probability matrix, $\mathbf{P}(v,\Theta_1) = \{p_{ij}(v,\Theta_1)\}$, specifies the probability that nucleotide i changes into j over a branch of length v. $\mathbf{P}(v,\Theta_1)$ can be obtained from the rate matrix **Q** as $\mathbf{P}(v,\Theta_1) = e^{\mathbf{Q}v}$.

The HKY85 model assumes that all sites in a DNA sequence have the same relative branch lengths (v). This assumption is usually referred to as the ''equal rates assumption,'' although it should be noted that the HKY85 model allows some rate variation across sites because the rate of substitution at a site depends on the current nucleotide at that position. Nonetheless, the equal rates assumption has been relaxed in several ways. For example, the sites of the sequence can be partitioned (e.g., into first, second, and third codon positions) and the rate in each partition estimated separately (see Swofford et al. 1996). Another commonly used method for accommodating among-site rate variation is to assume that the rate at each site (r) is a random variable drawn from some distribution (Felsenstein 1981). The most commonly used distributions assume that some proportion of the sites are invariant (Churchill et al. 1992; Hasegawa et al. 1985; Reeves 1992; Sidow et al. 1992; Waddell and Penny 1996) and/or that the rate at a site is drawn from a gamma distribution with shape parameter α (Gu et al. 1995; Jin and Nei 1990; Waddell and Penny 1996; Yang 1993).

In this study, we assume that the rate at each site is a random variable drawn from a gamma distribution. The gamma distribution is a continuous distribution with probability density

$$g(r\,|\alpha,\beta) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1}, & r \geqslant 0 \end{cases} \quad (2)$$

α is the shape parameter, and β is a scale parameter. The mean of the gamma distribution is $E(r) = \alpha/\beta$ and the variance is $\mathrm{Var}(r) = \alpha/\beta^2$. For the phylogeny problem, the scale parameter is set equal to the shape parameter ($\alpha_r = \beta_r$) so that the mean rate of substitution is 1 and the variance in the rate of substitution across sites is $\mathrm{Var}(r) = 1/\alpha_r$ (Jin and Nei 1990; Yang 1993) (the subscript ''r'' means that shape parameter is for the gamma distribution for rates across sites). For any particular realization of the rate at a site, the transition probability is calculated as $\mathbf{P}(v, r, \Theta_1) = e^{\mathbf{Q}vr}$. Hence, the effect of among-site rate variation is to modify the length of a branch; the substitution rate matrix **(Q)** is unaffected. The HKY85 model with gamma-distributed rate variation is denoted HKY85+Γ_r. The HKY85+Γ_r model of DNA substitution has free parameters $\Theta_2 = \{\kappa, \pi, \alpha_r\}$.

Although some models of DNA substitution are more complex than the HKY85 model because they include more substitution rate classes [e.g., the GTR model of Lanave et al. (1984) and Tavaré (1986)], all currently used models of DNA substitution are like the HKY85 model because they assume that the **Q** matrix is constant across sites. This is true whether or not the model accommodates among-site rate variation. We consider a more general case of the HKY85+Γ_r, model that allows the transition/transversion rate bias to vary across sites. We assume that the transition/transversion rate ratio at a particular site is modified by a gamma-distributed random variable with shape parameter α_κ. The instantaneous rate matrix for our model is

$$\mathbf{Q} = \{q_{ij}\} = r\mu \begin{pmatrix} \cdot & \pi_C & \bar{\kappa}y\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \bar{\kappa}y\pi_T \\ \bar{\kappa}y\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \bar{\kappa}y\pi_C & \pi_G & \cdot \end{pmatrix} \quad (3)$$

where all of the symbols are the same as in Eq. (1) but the transition/transversion rate parameter now represents the mean bias in the rate of transitions and transversions and is designated $\bar{\kappa}$. y represents the gamma-distributed random variable that modifies $\bar{\kappa}$ at a particular site in the sequence. For any realization of y, the rate matrix **Q** is rescaled such that the mean rate of substitution is 1. This model is designated HKY85+Γ_r+Γ_κ and includes parameters $\Theta_3 = \{\bar{\kappa}, \pi, \alpha_r, \alpha_\kappa\}$.

The gamma distribution was chosen to model variation in the transition/transversion rate ratio because of the wide range of shapes it can take. The gamma distribution has proven to be useful for modeling variation in the relative rates of substitution across sites (Yang 1996a). In fact, for among-site rate variation, the gamma distribution makes sense because for small values of the shape parameter, much of the weight of the gamma density is near zero rate and the density is zero for rates less than zero. However, it is not clear that the gamma distribution is ideal for modeling variation in the transition/transversion rate ratio. For example, it is usually thought that the transition/transversion rate ratio is greater than one. Hence, the use of the gamma distribution to

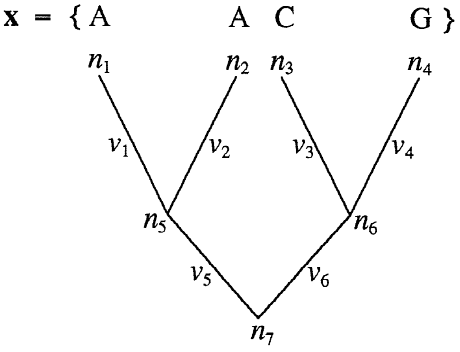

Fig. 1. We assume that the sequences are related by a phylogeny (τ). Here, the data at the tips of the tree are denoted $\mathbf{x} = \{A,A,C,G\}$. n_1, n_2, . . . , n_7 denote the nodes of the tree, with n_1, . . . , n_4 being external and n_5, n_6, and n_7 being internal nodes. The lengths of the branches are v_1, v_2, . . . , v_6.

model variation in the transition/transversion rate ratio should be considered as a starting point for further research. Other distributions, such as the gamma distribution without the constraint that $\alpha = \beta$ or distributions truncated such that the density is zero for transition/transversion rates less than one, may better describe variation in the pattern of DNA substitution.

The HKY85 and HKY85+Γ_r models of DNA substitution are simply special cases of the most complicated model (HKY85+Γ_r+Γ_κ) considered in this study. To obtain the HKY85 model from the HKY85+Γ_r+Γ_κ model, one sets $\alpha_r = \infty$ and $\alpha_\kappa = \infty$. Similarly, to obtain the HKY85+Γ_r model, one sets $\alpha_\kappa = \infty$. Because the models are nested, likelihood-ratio tests using the χ^2 approximation of the significance level can be performed comparing one model of DNA substitution to another (Goldman 1993). The ratio of the likelihoods obtained under null and alternative models ($\Lambda = L_0/L_1$) for the same data provides information about the relative fit of the models to the data. For the special case in which nested models are compared, the significance of the likelihood-ratio test statistic ($-2\log\Lambda$) can be compared to a χ^2 distribution with the appropriate degrees of freedom; the degree of freedom for the test is simply the difference in the number of free parameters between the null model and the more general alternative model. In this study, we are particularly interested in whether the addition of among-site variation in κ provides a significant improvement in the likelihood.

Calculation of the Likelihood. Calculation of the likelihood requires the specification of a model of evolution. For the phylogeny problem, the model includes not only a model of DNA substitution (see above) but also a tree topology (τ) with branch lengths specified in terms of the expected number of substitutions per site (Fig. 1). The branch lengths of the tree are contained in a vector $\mathbf{v} = \{v_1, v_2, \ldots, v_b\}$ (where $b = 2s - 3$ for unrooted trees and $b = 2s - 2$ for rooted trees). The likelihood is the probability of observing the data given a specified topology and the model of nucleotide substitution. Assuming independent substitution at sites, the likelihood for the sequences is

$$L(\tau,\mathbf{v},\Theta|\mathbf{X}) = \prod_{i=1}^{c} f(\mathbf{x}_i|\tau,\mathbf{v},\Theta) \tag{4}$$

Parameters in the model (τ, $\mathbf{v}$, and Θ) are estimated by maximizing the likelihood function.

For a given tree topology, the probablity of observing a site in the sequence ($\mathbf{x}$) is a sum over all possible assignments of nucleotide states to the internal nodes of the tree. For the tree shown in Fig. 1 and assuming independence of the random variables r and y, the probability of observing the data at the tips of the tree ($\mathbf{x} = \{A, A, C, G\}$) is

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\tau},\mathbf{v},\boldsymbol{\Theta}_3) &= \int_0^\infty\int_0^\infty f(\mathbf{x}|\boldsymbol{\tau},\mathbf{v},r,y,\Theta_3)g(r|\alpha_r,\alpha_r)g(y|\alpha_\kappa,\alpha_\kappa)dr\,dy \\ &= \int_0^\infty\int_0^\infty \Bigg\{\sum_{n5}\sum_{n6}\sum_{n7} \pi_{n_7} p_{n_7 n_5}(v_5,r,y,\Theta_3)p_{n_7 n_6} \\ &\quad \times (v_6,r,y,\Theta_3)p_{n_5 A}(v_1,r,y,\Theta_3)p_{n_5 A}(v_2,r,y,\Theta_3)p_{n_6 C}(v_3,r,y,\Theta_3) \\ &\quad \times p_{n_6 G}(v_4,r,y,\Theta_3)\Bigg\} g(r|\alpha_r,\alpha_r)g(y|\alpha_\kappa,\alpha_\kappa)dr\,dy \end{aligned} \tag{5}$$

The integrals are over all possible r and y for the site. This multidimensional integral was approximated by using the discrete gamma approach (Yang 1994). The gamma distributions for substitution rates and the κ multiplier were broken into $k = 5$ categories of equal frequency and the mean r or y from each category was used to represent the entire category (Yang 1994). An alternative way of approximating the integral is to use Monte Carlo integration (e.g., Fishman 1996). For the Monte Carlo approach, Eq. (5) is approximated by repeatedly simulating random variables from $g(r|\alpha_r, \alpha_r)$ and $g(y|\alpha_\kappa, \alpha_\kappa)$. If we denote the ith realization of r and y, r_i and y_i, respectively, then by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}|\boldsymbol{\tau},\mathbf{v},r_i,y_i,\Theta_3) \rightarrow f(\mathbf{x}|\boldsymbol{\tau},\mathbf{v},\boldsymbol{\Theta}_3) \tag{6}$$

as $n \rightarrow \infty$ and we can therefore obtain an unbiased estimate of Eq. (5) as

$$(1/n)\sum_{i=1}^{n} f(\mathbf{x}|\boldsymbol{\tau},\mathbf{v},r_i,y_i,\Theta_3),$$

where n is a large number. It was found empirically that 10^4 to 10^6 simulations for each site pattern provided reasonable estimates of the likelihood. The Monte Carlo approach was also used and provided similar results to the discrete gamma approximation. Only the results for the discrete gamma approximation are reported in this paper.

Data Analysis. We examined the pattern of nucleotide substitution for 14 genes. Thirteen of the genes are protein coding genes from the mitochondria of 10 vertebrates [ATPase 6, ATPase 8, cytochrome oxidase subunits I, II, and III (COI, COII, and COIII), cytochrome *b* (CYTB), NADH dehydrogenase subunits 1, 2, 3, 4, 5L, 5, and 6 (NADH1, NADH2, NADH3, NADH4, NADH4L, NADH5, and NADH6) (Cummings et al. 1995)]. One of the genes is a pseudogene ($\psi\eta$-globin) from six primates (Miyamoto et al. 1987). We used the alignments of Cummings et al. (1995) and Miyamoto et al. (1987) for the 13 mitochondrial genes and pseudogene, respectively.

Likelihoods were optimized on the trees shown in Fig. 2. The tree in Fig. 2A is the optimal tree based on the entire mitochondrial genome found in the Cummings et al. (1995) study. The tree in Fig. 2B, used for the pseudogene, represents the best estimate of phylogeny for apes and two monkeys currently available. Likelihoods for the HKY85 and HKY85+Γ_r models were optimized using a tester version of the program PAUP* 4.0 [version d57 (Swofford 1997)]. Maximum-likelihood estimates of parameters for each gene under the HKY85+Γ_r+Γ_κ model were calculated using programs written independently in C by both authors.

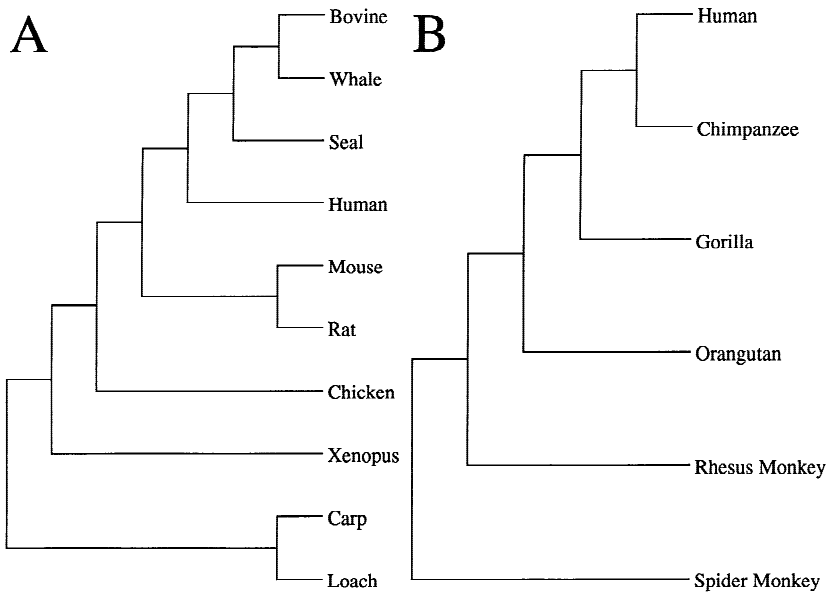

Fig. 2. The trees used for calculating likelihoods. **A** The tree assumed for the 13 mitochondrial genes. **B** The tree assumed for the pseudogene.

Results

Significant Variation in Transition/Transversion Rate Ratio Across Sites

We tested the assumption that the pattern of nucleotide substitution is constant across sites. Table 1 summarizes the analyses of the 13 protein coding genes and the pseudogene. For all genes examined, the rate of substitution varied significantly across sites (H_0, HKY85; H_1, HKY85+Γ_r). Estimates of the gamma shape parameter for rates ranged from $\alpha_r = 0.16$ to $\alpha_r = 0.68$ for protein coding genes and $\alpha_r = 1.22$ for the pseudogene. Adding gamma-distributed rate variation produced likelihood-ratio test statistics that ranged from $-2\log\Lambda = 110.84$ to $-2\log\Lambda = 2055.38$ for coding genes and $-2\log\Lambda = 18.44$ for the pseudogene. The 5% significance level for a χ^2 distribution with 1 df is $\chi^2_{(1,0.05)} = 3.84$, so all values for the likelihood-ratio test statistic are much greater than would be expected if the null hypothesis were true. Moreover, for 12 of the 13 coding genes the null hypothesis of a constant transition/transversion rate ratio across sites can be rejected (H_0, HKY85+Γ_r; H_1, HKY85+Γ_r+Γ_κ; the likelihood-ratio test statistic varied from $-2\log\Lambda = 5.04$ to $-2\log\Lambda = 201.34$). The null hypothesis could not be rejected for NADH6 ($-2\log\Lambda = 1.86$), which was among the smallest coding genes considered. Furthermore, the null hypothesis of a constant transition/transversion rate ratio across sites could not be rejected for the pseudogene ($-2\log\Lambda = 0.0$). It is unlikely that failure to reject the null hypothesis for the pseudogene results from lack of data because the pseudogene was 3.3 times longer than the longest coding gene considered.

Although the effect of adding among-site variation in the transition/transversion rate ratio provides a significant improvement in the fit of the model to the data, the improvement in fit is small when compared to the improvement made when adding rate variation across sites. Tables 1 and 2 also include log likelihoods calculated under the HKY85+Γ_κ model of DNA substitution. Note that the improvement in fit is several hundred for the HKY85+Γ_r model of DNA substitution, whereas the improvement is typically much smaller for the HKY85+Γ_κ model.

One possible explanation for the significant variation in the transition/transversion rate ratio (κ) across sites is that κ is different among codon positions but constant among sites assigned to a codon. The proportion of substitutions that are transitions is higher among potential synonymous substitutions than nonsynonymous substitutions. In twofold degenerate sites, all potential synonymous substitutions are transitions, whereas all potential nonsynonymous substitutions are transversions. Under the assumption that the rate of nonsynonymous substitution is lower than the rate of synonymous substitution, the transition/transversion ratio will vary among codon positions simply because of the nature of the genetic code. This effect will be of importance even under very simple models of sequence evolution. For example, assume that the instantaneous rate of change in any position of the nucleotide sequence is given by $\mathbf{Q}^* = \{q_{ij}R_{ij}\}$, where q_{ij} is the instantaneous rate of change from nucleotide i to nucleotide j under the HKY85 model and where R_{ij} is 1 if the substitution is nonsynonymous and ω otherwise. Then ω is the rate ratio of nonsynonymous-to-synonymous substitutions and the model describes nucleotide evolution by the HKY85 model with selection. More applicable models of codon evolution have been considered by Muse and Gaut (1994) and Goldman and Yang (1994). Let ρ denote the relative increase in the ratio of transitions to transversions due to selection. The

Table 1. Models and parameter estimates obtained in analyses of 13 protein coding genes and one pseudogene

		HKY85		HKY85 + Γ_κ			HKY85 + Γ_r			HKY85 + Γ_r + Γ_κ			
Gene	c	log L	κ	log L	$\bar{\kappa}$	α_κ	log L	κ	a_r	log L	$\bar{\kappa}$	α_κ	α_r
ATPase6	687	−5,525.77	2.08	−5,509.03	3.49	0.83	−5,192.90	4.43	0.35	−5,172.64	11.3	0.47	0.32
ATPase8	207	−1,650.54	2.84	−1,649.44	3.62	2.02	−1,595.12	5.32	0.64	−1,592.60	8.94	0.89	0.60
COI	1,560	−10,606.41	2.41	−10,496.30	10.1	0.37	−9,578.72	8.38	0.16	−9,478.05	97.1	0.18	0.15
COII	705	−5,065.91	2.53	−5,027.42	7.49	0.48	−4,740.70	5.35	0.30	−4,694.54	37.1	0.24	0.27
COIII	785	−5,407.81	2.39	−5,362.35	9.37	0.37	−4,986.94	5.43	0.24	−4,930.05	47.8	0.21	0.22
CYTB	1,149	−8,201.93	2.05	−8,126.30	7.52	0.36	−7,684.42	3.49	0.34	−7,605.58	22.5	0.23	0.31
NADH1	981	−7,382.19	2.00	−7,325.61	5.65	0.42	−6,905.17	4.41	0.31	−6,845.67	23.9	0.24	0.28
NADH2	1,047	−9,354.33	1.74	−9,331.39	2.70	0.82	−8,967.78	2.99	0.57	−8,937.23	6.55	0.47	0.51
NADH3	350	−2,847.28	2.45	−2,845.08	4.21	0.93	−2,684.32	4.99	0.38	−2,674.51	11.7	0.53	0.36
NADH4	1,387	−11,347.37	1.89	−11,311.96	3.20	0.75	−10,739.72	3.44	0.42	−10,688.57	9.78	0.39	0.37
NADH4L	297	−2,498.08	1.70	−2,492.60	2.74	0.86	−2,424.43	2.43	0.68	−2,418.89	4.98	0.60	0.64
NADH5	1,860	−15,492.33	1.95	−15,447.68	3.11	0.82	−14,696.65	3.47	0.47	−14,646.44	7.25	0.51	0.43
NADH6	561	−4,545.74	2.70	−4,545.74	2.70	∞	−4,385.10	4.75	0.63	−4,384.17	5.86	2.29	0.60
ψη-Globin	6,166	−13,833.92	4.79	−13,833.92	4.79	∞	−13,824.70	4.93	1.22	−13,824.70	4.93	∞	1.22

difference between $\kappa\rho$ and κ will approximately be the bias, due to the effect of selection, in the estimate of κ in models that do not take account of selection and the structure of the genetic code. For a given set of sequences, ρ can be calculated as

$$\rho = \frac{\sum_{i=1}^{c} \sum_{j=a,c,t,g;\, j\neq b_i} q_{b_i j}\omega T_{b_i j}}{\sum_{i=1}^{c} \sum_{j=a,c,t,g;\, j\neq b_i} q_{b_i j}\omega(1 - T_{b_i j})} \Bigg/ \frac{\sum_{i=1}^{c} \sum_{j=a,c,t,g;\, j\neq b_i} q_{b_i j} T_{b_i j}}{\sum_{i=1}^{c} \sum_{j=a,c,t,g;\, j\neq b_i} q_{b_i j}(1 - T_{b_i j})},$$

$$T_{ij} = \begin{cases} 1 & i \rightarrow j \text{ is a transition} \\ 0 & \text{else} \end{cases} \qquad (7)$$

where c is the total length of the sequences, b_j is the nucleotide at position j, and ω is the selection factor (the rate ratio of nonsynonymous to synonymous substitutions). As an example, ρ was calculated separately for each codon position for the COII gene (Fig. 3). Note that ρ will be elavated for the first and, especially, the third positions. In the second position, however, uniformly acting selection has no effect on the transition/transversion bias because all potential substitutions are nonsynonymous. The observed variation in ρ among codon positions may explain a large proportion of the apparent variation in κ among sites.

To test for any residual variation in κ when variation among codon positions has been accounted for, we also estimated parameters of the HKY85+Γ_r+Γ_κ model for the protein coding genes separately for the first, second, and third codon positions. Table 2 summarizes the results of these analyses. As expected from the above analysis, the estimate of κ varied considerably among positions. However, considerable residual variation in κ within codon positions was also observed. In general, the variation in the transition/transversion rate ratio was strongest for the first and second positions but weakest in the third position. The null hypothesis of no variation in κ across sites could not be rejected for the first codon position for 3 genes (ATPase8, NADH3, and NADH6), could not be rejected for the second codon position for 5 genes (ATPase6, ATPase8, NADH2, NADH4L, and NADH6), and could not be rejected for the third codon position for 10 genes (ATPase8, COI, COII, COIII, CTYB, NADH1, NADH3, NADH4L, NADH5, and NADH6).

Estimating Parameters When the Substitution Pattern Varies Across Sites

We examined the effect on parameter estimation of variation in the transition/transversion rate ratio across sites by generating sequences under the model of Kimura (1980; designated K80) with gamma-distributed transition/transversion rate bias (K80+Γ_κ). We estimated genetic distances and the transition/transversion rate ratio using the K80 model of DNA substitution [with κ and d, the genetic distance, estimated for pairwise comparisons using the method of Jin and Nei (1990)]. The K80 model of DNA substitution is a special case of the more general HKY85 model of DNA substitution and allows for a transition/transversion rate bias but assumes that all base frequencies are equal. It also assumes that the transition/transversion rate ratio is constant across sites; hence, the assumptions of the K80 model are violated because the sequences were generated with variation in the transition/transversion ratio across sites. For pairwise comparisons of sequences, there are only 16 possible site patterns that can occur (AA, AC, AG, AT, CA, CC, . . . , TT). For any combination of genetic distance between sequences (d, in terms of expected number of substitutions per site)

Table 2. Models and parameter estimates obtained for the protein coding genes partitioned by codon

		HKY85		HKY85 + Γ_κ			HKY85 + Γ_r			HKY85 + Γ_r + Γ_κ			
Gene	c	log L	κ	log L	$\bar{\kappa}$	α_κ	log L	κ	α_r	log L	$\bar{\kappa}$	α_κ	α_r
ATPase6													
1st	229	−1607.34	3.10	−1602.05	5.45	0.86	−1540.45	5.00	0.48	−1528.35	17.3	0.42	0.39
2nd	229	−953.10	2.11	−951.78	3.78	0.89	−918.74	2.89	0.32	−917.48	4.94	1.01	0.31
3rd	229	−2446.57	4.44	−2424.03	19.7	0.37	−2416.85	23.2	0.72	−2412.09	34.2	0.42	1.04
ATPase8													
1st	69	−525.22	3.27	−525.22	3.27	∞	−513.24	4.66	0.80	−513.24	4.66	∞	0.80
2nd	69	−441.84	1.69	−439.79	5.36	0.50	−431.79	2.09	0.70	−430.14	8.22	0.40	0.69
3rd	69	−621.80	5.18	−618.62	11.5	0.56	−609.03	14.3	0.78	−608.93	18.4	1.21	0.78
COI													
1st	520	−2188.35	4.21	−2180.40	11.4	0.68	−2067.00	5.32	0.21	−2054.42	22.2	0.45	0.20
2nd	520	−1147.61	1.69	−1146.07	4.49	0.50	−1124.06	1.67	0.01	−1122.01	9.66	0.27	0.01
3rd	520	−5542.72	6.17	−5397.60	100	0.23	−5384.01	498	0.29	−5384.01	498	∞	0.29
COII													
1st	235	−1363.84	2.70	−1361.42	4.43	0.97	−1315.86	3.51	0.46	−1309.72	9.14	0.48	0.42
2nd	235	−696.86	2.93	−691.66	100	0.19	−682.15	3.21	0.26	−677.85	92.3	0.20	0.27
3rd	235	−2411.56	5.86	−2363.33	56.3	0.27	−2353.72	170	0.39	−2353.72	170	∞	0.39
COIII													
1st	187	−970.90	3.42	−967.28	8.72	0.64	−925.26	4.52	0.31	−917.68	24.7	0.35	0.27
2nd	187	−560.73	3.94	−553.61	100	0.23	−545.33	4.19	0.21	−538.48	>600	0.17	0.22
3rd	187	−1911.88	6.02	−1878.06	50.7	0.30	−1869.05	93.8	0.47	−1869.05	93.8	∞	0.47
CYTB													
1st	383	−2259.53	2.60	−2251.06	5.44	0.73	−2125.72	3.61	0.30	−2109.16	15.7	0.34	0.27
2nd	383	−1263.84	3.58	−1259.54	15.7	0.45	−1223.04	3.78	0.23	−1219.32	16.1	0.46	0.23
3rd	383	−3746.32	6.19	−3664.57	52.3	0.27	−3671.49	147	0.46	−3671.48	149	12.6	0.47
NADH1													
1st	327	−2096.30	2.90	−2091.97	4.70	1.09	−2002.92	4.16	0.43	−1992.82	11.1	0.53	0.37
2nd	327	−1203.22	2.23	−1197.77	8.91	0.37	−1134.62	2.39	0.18	−1131.70	7.95	0.38	0.17
3rd	327	−3299.02	6.46	−3225.48	100	0.22	−3219.34	289	0.35	−3218.62	389	0.31	0.49
NADH2													
1st	349	−3021.27	1.45	−3011.96	2.13	0.68	−2928.43	1.91	0.77	−2914.09	4.06	0.38	0.67
2nd	349	−2008.88	2.48	−2005.00	4.61	0.99	−1954.27	2.70	0.60	−1952.76	4.22	1.30	0.60
3rd	349	−3706.44	4.78	−3663.28	28.0	0.31	−3656.29	24.9	0.68	−3651.44	39.5	0.41	0.89
NADH3													
1st	117	−863.79	2.66	−863.87	2.66	∞	−819.78	4.11	0.45	−819.66	4.69	3.11	0.43
2nd	117	−519.20	5.31	−515.96	35.2	0.43	−490.49	7.51	0.22	−487.67	64.3	0.37	0.20
3rd	116	−1195.02	6.40	−1178.47	39.1	0.32	−1177.89	63.6	0.62	−1177.89	63.6	∞	0.62
NADH4													
1st	463	−3334.44	2.17	−3330.09	3.05	1.22	−3217.77	3.12	0.58	−3204.38	7.82	0.48	0.49
2nd	462	−2150.66	1.74	−2141.86	4.57	0.53	−2056.31	2.06	0.31	−2050.36	5.29	0.52	0.30
3rd	462	−4895.19	5.50	−4834.48	29.6	0.35	−4830.09	34.8	0.68	−4825.17	33.9	0.47	1.25
NADH4L													
1st	99	−793.23	1.80	−791.15	2.66	0.83	−779.06	2.15	1.02	−775.11	5.40	0.38	0.77
2nd	99	−463.21	2.58	−463.21	2.66	17.7	−456.63	2.91	0.80	−456.39	4.56	1.35	0.73
3rd	99	−1037.43	4.03	−1026.15	17.1	0.36	−1023.28	23.9	0.67	−1023.30	23.3	7.48	0.70
NADH5													
1st	620	−4719.38	1.61	−4712.71	2.14	1.13	−4496.38	2.35	0.50	−4484.30	4.13	0.58	0.46
2nd	620	−3302.81	1.93	−3291.23	4.06	0.62	−3169.94	2.27	0.41	−3163.91	4.19	0.70	0.40
3rd	620	−6413.76	8.23	−6338.94	34.0	0.44	−6339.18	69.8	0.74	−6339.18	69.8	∞	0.74
NADH6													
1st	187	−1355.65	2.81	−1355.65	2.81	∞	−1323.81	4.59	0.76	−1323.68	5.37	3.48	0.73
2nd	187	−1035.04	1.72	−1035.04	1.72	∞	−1029.38	1.84	1.90	−1029.38	1.84	∞	1.90
3rd	187	−1828.31	132	−1821.12	61.8	0.94	−1819.59	88.9	0.91	−1819.59	88.9	∞	0.91

and transition/transversion rate ratio (κ), we calculated the expected frequencies of each of the 16 site patterns. These frequencies were then used in the analysis. We set $\bar{\kappa} = 5$ in our analyses.

Failure to accommodate variation in the pattern of DNA substitution biases estimates of evolutionary distance between species. Figure 4A shows the estimated distance ($\hat{d}$) as a function of the true distance (d) between species. The distance between the sequences is underestimated and the underestimate becomes more severe as the variation in the pattern of DNA substitution increases. Variation in the pattern of substitution also effects estimation of other parameters of interest to evolutionary biologists. Figure 4B shows the estimate of the

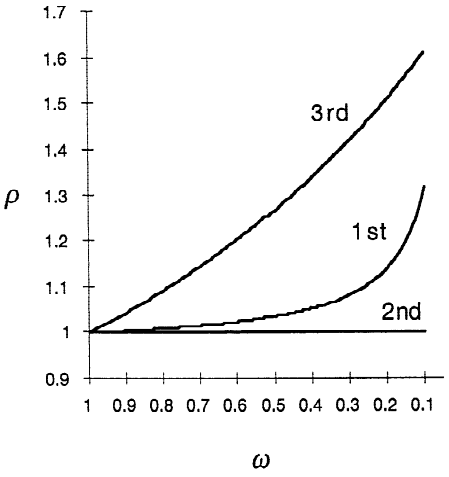

Fig. 3. The effect of selection on the transition/transversion bias in the first, second, and third positions of the codon, calculated for the COII gene. ρ is the relative increase in the transition/transversion bias and ω is the ratio of nonsynonymous to synonymous substitutions.

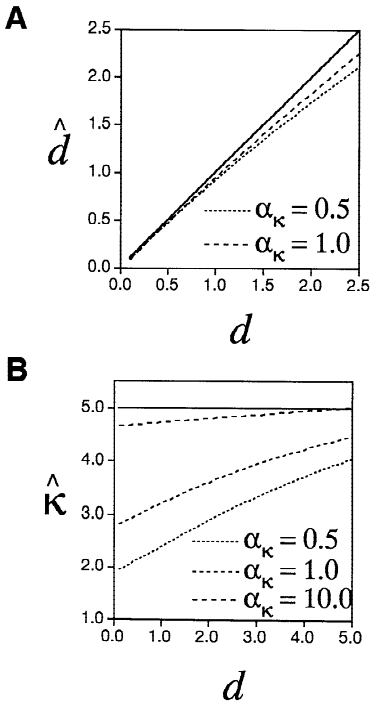

Fig. 4. The bias introduced to the estimation of evolutionary parameters when variation in the pattern of nucleotide substitution is not accommodated. Sequences were generated under a K80 model of substitution with $\kappa = 5$ multiplied by a random variable with shape parameter α_K. **A** The expected bias in the estimate of distance ($\hat{d}$) between sequences. **B** The expected bias in estimating the average transition/transversion rate ratio ($\hat{\kappa}$).

average transition/transversion rate ratio ($\hat{\kappa}$) as a function of the true distance between species. The average transition/transversion rate ratio is consistently underestimated.

Examination of the parameter estimates from the real data confirms the intuition provided by the theoretical analysis. In almost every instance, the estimate of κ for the HKY85+Γ_r+Γ_κ model is equal to or greater than the estimate obtained under the HKY85+Γ_r model of DNA substitution. In some cases, the underestimate of κ can be quite severe. For example, the estimate of κ changed from 8.4 to 20.9 for COI when among-site variation in the transition/transversion bias was accommodated.

Discussion

We have shown that substantial variation in the ratio of transitions to transversion occurs among sites in coding genes. Purifying selection acting uniformly in the genes may explain this pattern because of intrinsic properties of the genetic code. However, it is interesting to note that κ also appears to vary within codon positions. Especially, the observed variation of κ in the second position cannot be explained simply by the structure of the genetic code because all changes in the second position are nonsynonymous. This suggests that some residual variation in κ is caused either by variation in the mutational process or by other effects of selection. Because the residual effects appear to be larger in the first and second position than in the third position of the codon, and because no significant variation in κ was observed in the pseudogene, selection appears at the present to be the most plausible explanation.

Regardless of the explanation, the variation in the pattern of substitution across sites is substantial and may need to be accommodated by phylogenetic methods for accurate inference of evolutionary parameters. The underestimation of genetic distances (d) between sequences and the transition/transversion rate ratio (κ) can be large. This result is similar to the bias introduced into estimates of d and κ when among-site rate variation is not accommodated in a phylogenetic analysis (Wakeley 1994; Yang 1996a). In a likelihood framework, accommodating variation in κ across sites can be accomplished by using the discrete gamma approximation adapted from Yang (1994). However, use of the gamma approximation both for the substitution rate and for the transition bias slows calculation of likelihoods. In our analyses, we used five categories to approximate the gamma distributions for rates and transition bias. This meant that the likelihood for a site was a sum over $5^2 = 25$ categories. Hence, analyses were five times slower than they would have been if only rate variation among sites had been accommodated.

Acknowledgments. We thank Monty Slatkin for helpful advice during the course of this study. David Swofford and Andrew MacArthur made valuable comments on an early version of the manuscript. The comments of an anonymous reviewer greatly improved the manuscript. This research was supported by an NSF/Sloan postdoctoral fellowship in molecular evolution awarded to J.P.H., a NIH grant awarded to Montgomery Slatkin, and a fellowship awarded to R.N. from the Danish Research Council.

References

Churchill GA, von Haeseler A, Navidi WC (1992) Sample size for a phylogenetic inference. Mol Biol Evol 9:753–769

Cummings MP, Otto SP, Wakeley J (1995) Sampling properties of DNA sequence data in phylogenetic analysis. Mol Biol Evol 12: 814–822

Farris JS (1969) A successive approximations approach to character weighting. Syst Zool 18:374–385

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368–376

Fishman GS (1996) Monte Carlo. Springer-Verlag, New York

Golding GB (1983) Estimates of DNA and protein sequence divergence: An examination of some assumptions. Mol Biol Evol 1:125–142

Goldman N (1993) Statistical tests of models of DNA substitution. J Mol Evol 36:182–198

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736

Goloboff PA (1993) Estimating character weights during tree searches. Cladistics 9:83–91

Gu X, Fu Y-X, Li W-H (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 12:546–557

Hasegawa M, Yano T, Kishino H (1984) A new molecular clock of mitochondrial DNA and the evolution of Hominoids. Proc Japan Acad Ser B 60:95–98

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape split by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogeny analysis. Mol Biol Evol 7:82–102

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

Miyamoto MM, Slightom JL, Goodman M (1987) Phylogenetic relationships of humans and African apes as ascertained from DNA sequences (7.1 kilobase pairs) of the ψη-globin region. Science 238:369–373

Muse SV, Gaut S (1994) A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

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

Sidow A, Nguyen T, Speed TP (1992) Estimating the fraction of invariable codons with a capture-recapture method. J Mol Evol 35: 253–260

Swofford DL, Olsen GP, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics, 2nd ed. Sinauer, Sunderland, MA, pp 407–514

Tavaré S (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. Lect Math Life Sci 17:57–86

Waddell PJ, Penny D (1996) Evolutionary trees of apes and humans from DNA sequences. In: Lock AJ, Peters CR (eds) Handbook of symbolic evolution. Clarendon Press, Oxford, pp 53–73

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

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 (1996a) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372

Yang Z (1996b) Maximum likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587–596

Yang Z, Goldman N, Friday AE (1994) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 11:316–324