JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1998

Synonymous and Nonsynonymous Rate Variation in Nuclear Genes of Mammals

Ziheng Yang,[1,2,]* Rasmus Nielsen[1,]**

[1] Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA
[2] Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, England

Received: 17 May 1997 / Accepted: 28 September 1997

Abstract. A maximum likelihood approach was used to estimate the synonymous and nonsynonymous substitution rates in 48 nuclear genes from primates, artiodactyls, and rodents. A codon-substitution model was assumed, which accounts for the genetic code structure, transition/transversion bias, and base frequency biases at codon positions. Likelihood ratio tests were applied to test the constancy of nonsynonymous to synonymous rate ratios among branches (evolutionary lineages). It is found that at 22 of the 48 nuclear loci examined, the nonsynonymous/synonymous rate ratio varies significantly across branches of the tree. The result provides strong evidence against a strictly neutral model of molecular evolution. Our likelihood estimates of synonymous and nonsynonymous rates differ considerably from previous results obtained from approximate pairwise sequence comparisons. The differences between the methods are explored by detailed analyses of data from several genes. Transition/transversion rate bias and codon frequency biases are found to have significant effects on the estimation of synonymous and nonsynonymous rates, and approximate methods do not adequately account for those factors. The likelihood approach is preferable, even for pairwise sequence comparison, because more-realistic models about the mutation and substitution processes can be incorporated in the analysis.

Key words: Synonymous rates — Nonsynonymous rates — Mammalian genes — Likelihood — Codon substitution — Transition/transversion rate bias — Neutral theory

Introduction

Estimation of synonymous and nonsynonymous substitution rates is important in understanding the dynamics of molecular sequence evolution (Kimura 1983; Gillespie 1991; Ina 1996). As mutation and selection have different effects on synonymous and nonsynonymous substitutions, comparison of synonymous and nonsynonymous rates in genes from different evolutionary lineages at different loci provides a powerful tool for understanding the mechanisms and driving forces of molecular evolution. For example, the neutral theory (Kimura 1983) claims that both divergence between species and diversity within species are caused by random genetic drift of neutral mutations (Kimura 1968) and the substitution rate is equal to the neutral mutation rate. The effect of purifying selection is to reduce the proportion of neutral mutations. One prediction of the theory is that the nonsynonymous/synonymous substitution rate ratio should be constant among different evolutionary lineages. Lineage effects of mutation rates should have the same effect on synonymous and nonsynonymous rates

* *Present address:* Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, England; e-mail z.yang@ucl.ac.uk
** *Present address:* The Museum of Comparative Zoology, 3rd floor, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA
Correspondence to: Z. Yang

and not change the rate ratio. Variation of the nonsynonymous/synonymous rate ratios among lineages is thus considered evidence against neutrality (McDonald and Kreitman 1991; Gillespie 1991; Easteal and Collet 1994; Eyre-Walker and Gaut 1997).

Gillespie (1987, 1989, 1991) and Ohta (1995) estimated the index of dispersion for synonymous and nonsynonymous substitutions in nuclear genes of mammals to test the neutral theory. The index of dispersion is the variance of the number of substitutions among lineages divided by the mean. Under neutrality, the substitution process should be approximately Poisson and the index of dispersion should equal one. However, Gillespie's (1987, 1989) and Ohta's (1995) estimates of the index are much greater than one. To explain the discrepancy, Gillespie (1991) suggested that nonsynonymous substitutions were driven by positive selection, resulting in episodic evolution with bursts of substitutions, while Ohta (1995) invoked the fixation of slightly deleterious mutations. However, the methods used by those authors are crude, and it has been suggested that their high estimates of the dispersion index may be artifacts of the analytical methods used (Goldman 1994; Nielsen 1997). It is not clear for how many of the examined genes the neutral model of evolution can be rejected in favor of positive selection or selection on slightly deleterious mutations.

Although its importance is well recognized, estimation of synonymous and nonsynonymous substitution rates is not simple. Important concepts were developed in the 1980s for estimating the rates from a comparison of two sequences (Miyata and Yasunaga 1980; Li et al. 1985), and the early methods have been improved or simplified by many authors (Nei and Gojobori 1986; Li 1993; Pamilo and Bianchi 1993; Cameron 1995). Those methods follow the same strategy. The numbers of synonymous (S) and nonsynonymous (N) sites in the sequence and the numbers of synonymous (S_d) and nonsynonymous (N_d) differences between the two sequences are counted. Corrections for multiple substitutions are then applied to calculate the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site between the two sequences (see Table 1 for definitions of symbols). These methods assume equal base and codon frequencies and do not account properly for the transition/transversion rate bias in counting the numbers of sites (S and N) and differences (S_d and N_d). The method recently proposed by Ina (1995) is a significant improvement and accounts for the transition/transversion bias. A drawback of the approximate methods is that they cannot be used for simultaneous comparison of multiple sequences.

The codon-based likelihood model suggested by Goldman and Yang (1994) provides a useful framework for estimating synonymous and nonsynonymous substitution rates. Under a model of codon substitution, it is straightforward to estimate d_S and d_N (Goldman and Yang 1994). Knowledge of the substitution process such as transition/transversion rate bias, codon frequency biases, and even amino acid differences can be easily incorporated into the model. Furthermore, the likelihood approach is applicable to joint comparison of multiple sequences (Goldman and Yang 1994).

Table 1. Definitions of major symbols used in the paper

Symbol	Definition
S	Number of synonymous sites in a sequence
N	Number of nonsynonymous sites in a sequence
S_d	Number of synonymous differences between two sequences
N_d	Number of nonsynonymous differences between two sequences
d_S	Number of synonymous substitutions per synonymous site
d_N	Number of nonsynonymous substitutions per nonsynonymous site
κ	Transition/transversion (mutation) rate ratio
ω	Nonsynonymous/synonymous rate ratio, equal to d_N/d_S under the model of Eq. 1
t	Time or branch length, measured as the expected number of (nucleotide) substitutions per codon
π_j	Equilibrium frequency of codon j
L_c	Number of codons in the sequence

The aim of this study is to reanalyze Ohta's (1995) data by a likelihood approach to examine the claim of Gillespie and Ohta that amino acid replacements in the analyzed genes are not neutral. Ohta (1995) used Ina's (1995) method for pairwise sequence comparison to estimate the numbers of synonymous and nonsynonymous substitutions among the primate, artiodactyl, and rodent lineages in 49 nuclear genes. We reanalyze Ohta's data using a likelihood model that accounts for different d_N/d_S ratios among different lineages and another model that constrains the d_N/d_S ratio to be constant. The two models are compared by a likelihood ratio test to examine the null hypothesis that the d_N/d_S ratio is constant among lineages. The analysis is performed independently for each locus and may help to determine for how many and which of the examined genes neutrality can be rejected. Our estimates of the synonymous and nonsynonymous rates turn out to differ considerably from those of Ohta (1995). We therefore compare different methods for rate estimation to understand the observed differences.

Data and Methods

Sequence Data. The aligned sequences of 49 nuclear genes from primates, artiodactyls, and rodents analyzed by Ohta (1995) were used. The opsin gene alignment has many gaps, indicating that the alignment may not be reliable, and this gene is excluded. Data used in this study include 48 genes from the three orders of mammals, with a total of 18,630 x 3 = 55,890 nucleotide sites. The number of codons in each gene used is shown in Table 2. Minor differences in sequence length from Ohta (1995: Table 3) are due to our removal of the initiation codons (ATG) and minor adjustments to Ohta's alignments.

Model of Codon Substitution. We use a simplified version of the codon substitution model developed by Goldman and Yang (1994). Stop codons are not allowed in the sequence, and substitutions between sense codons are described by a continuous-time Markov process. The instantaneous substitution rate from codon i to j ($i \neq j$) is given by

$$Q_{ij} = \begin{cases} 0 & \text{if the two codons differ at more} \\ & \quad \text{than one position} \\ \mu\pi_j & \text{for synonymous transversion} \\ \mu\kappa\pi_j & \text{for synonymous transition} \\ \mu\omega\pi_j & \text{for nonsynonymous transversion} \\ \mu\omega\kappa\pi_j & \text{for nonsynonymous transition} \end{cases} \tag{1}$$

where parameter κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio, and π_j is the equilibrium frequency of codon j. The diagonals of the matrix are determined by the mathematical requirement that row sums of $Q = \{Q_{ij}\}$ are zero. The scale factor μ is chosen such that

$$-\sum_i \pi_i Q_{ii} = \sum_i \pi_i \sum_{j \neq i} Q_{ij} = 1 \tag{2}$$

This scaling means that time and branch length (denoted t) are measured by the expected number of (nucleotide) substitutions per codon.

Goldman and Yang (1994) used the matrix of amino acid distances of Grantham (1974) to modify nonsynonymous substitution rates, based on the expectation that amino acids with similar physicochemical properties tend to replace each other more often than those with different properties (e.g., Miyata and Yasunaga 1980; Li et al. 1985). However, the formula used was found to fit real data poorly (Goldman and Yang 1994), and in this study, we make no distinction among different amino acid changes. The model is equivalent to that of Goldman and Yang using a single distance between any pair of amino acids. The model of Muse and Gaut (1994) differs from the present model in that those authors did not account for the transition/transversion rate bias.

Calculation of the likelihood function under the codon-substitution model was described by Goldman and Yang (1994) and Muse and Gaut (1994). The transition probability matrix, $P(t) = e^{Qt}$, is calculated through diagonalization of the rate matrix Q, with a standard numerical algorithm used to calculate the eigenvalues and eigenvectors of Q. Numerical optimization algorithms are used to obtain maximum likelihood estimates of parameters.

Test for Constant Nonsynonymous/Synonymous Rate Ratios (d_N/d_s) *Among Lineages.* The codon substitution model (Eq. 1) can be used to construct various tests concerning the evolutionary process of protein-coding DNA sequences. In this study, we fit a model assuming different nonsynonymous/synonymous rate ratios ($d_N/d_S = \omega$; see below) among branches in the phylogenetic tree and another model that constrains the ratio to be identical. These two models are depicted in Fig. 1, using the example of the data analyzed in this study. Comparison of the two models constitutes a likelihood ratio test of the constancy of the d_N/d_s ratio among evolutionary lineages.

Estimation of Synonymous and Nonsynonymous Substition Rates. After maximum likelihood estimates of model parameters (such as t, κ, and ω) are obtained, the numbers of synonymous and nonsynonymous substitutions per site (d_S and d_N) can be estimated as follows (Goldman and Yang 1994). Since the Q matrix is scaled such that the average number of nucleotide substitutions per codon is one,

$$\rho^*_S = \sum_{\substack{i \neq j \\ aa_i = aa_j}} \pi_i Q_{ij} \tag{3}$$

and $\rho^*_N = 1 - \rho^*_S$ are the proportions of synonymous and nonsynonymous substitutions, respectively. The summation in Eq. 3 is taken over all codon pairs i and j ($i \neq j$) that code for the same amino acid, and aa_i is the amino acid encoded by codon i. The numbers of synonymous and nonsynonymous substitutions per codon are then $t\rho^*_S$ and $t\rho^*_N$, respectively. The proportions of synonymous and nonsynonymous sites are defined as the proportions of synonymous and nonsynonymous potential mutations before natural selection at the amino acid level has operated (Goldman and Yang 1994; Ina 1995). Denote them as ρ^1_S and ρ^1_N (equivalent to ρ^∞_S and ρ^∞_N in Goldman and Yang 1994). They can be calculated similarly to Eq. 3 but with ω set at 1 so that synonymous and nonsynonymous substitutions occur at the same rate. As each codon has three sites (but see Discussions), the numbers of synonymous and nonsynonymous sites per codon are $3\rho^1_S$ and $3\pi^1_N$, respectively. The numbers of synonymous and nonsynonymous substitutions per site are then $d_S = t\rho^*_S/3\rho^1_S$ and $d_N = t\rho^*_N/3\rho^1_N$, respectively. Note that under the model of Eq. 1, $d_N/d_S = \omega$.

Approximate Methods for Estimating Synonymous and Nonsynonymous Rates. Approximate methods for estimating the numbers of synonymous and nonsynonymous substitutions between a pair of sequences were reviewed by Ina (1996). We used the methods of Nei and Gojobori (1986) and Ina (1995) to perform pairwise comparisons of sequences from several genes and compare the results with those of the likelihood analysis assuming different models. In fact, two methods were proposed by Ina (1995). Method 1 estimates the transition/transversion rate ratio (κ) from data at the third codon positions, while method 2 uses a more-sophisticated iterative scheme to estimate κ. Ohta (1995) used Ina's method 1 and called it Ina's method, a practice we have followed in this paper. Ina's program (new1, available from ftp.nig.ac.jp) is used in the calculation. The PAML package (Yang 1997) is used for Nei and Gojobori's method.

Results

Test for Constancy of Nonsynonymous/Synonymous Rate Ratios (d_N/d_S) *Among Lineages*

Maximum likelihood estimation was carried out for each gene, under model A (Fig. 1), assuming the same d_N/d_S ratio (ω) for all branches in the tree, and under model B (Fig. 1), assuming different ratios for different branches. Under both models, the codon frequencies (π_i) were calculated using the nucleotide frequencies at the three codon positions; 3 × (4 − 1) = 9 parameters are thus used for the codon frequencies. The results are presented in Table 2. Estimates of κ under the two models are almost identical for each gene and range from 2 to 5 among genes with a mean 2.9. Estimates of ω ($= d_N/d_S$) shown in Table 2 are obtained under model A and should be interpreted as averages across the three branches. They range from 0.017 for the highly conserved ATP synthase β to 0.838 for interleukin 6, with Mean ± S.E. to be 0.17 ± 0.03. On average, synonymous substitutions occur much more often than nonsynonymous substitutions, in concordance with earlier studies (Miyata and Yasunaga 1980; Li et al. 1985).

Table 2. Maximum likelihood estimates of the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site[a]

Gene	L_c	κ	ω	d_S (P)	d_S (A)	d_S (R)	d_N (P)	d_N (A)	d_N (R)	Δℓ
Acetylcholine receptor α	456	3.048	0.052	0.147	0.165	0.404	0.013	0.007	0.018	1.22
Acetylcholine receptor β	500	3.061	0.099	0.204	0.187	0.412	0.023	0.020	0.036	0.30
Acid phosphatase type 5	322	4.181	0.098	0.354	0.259	0.680	0.028	0.049	0.049	3.11[b]
Albumin	606	1.962	0.229	0.291	0.160	0.709	0.055	0.094	0.106	12.26[c]
Alkaline phosphatase intestine	495	2.006	0.174	0.343	0.328	0.534	0.041	0.088	0.081	3.49[b]
Alkaline phosphatase liver	523	2.499	0.074	0.259	0.340	0.650	0.025	0.025	0.041	0.84
Aspartate aminotransferase cytosolic	412	2.489	0.095	0.147	0.230	0.380	0.016	0.026	0.028	0.84
Aspartate aminotransferase mitochondrial	429	2.827	0.063	0.158	0.233	0.362	0.013	0.023	0.011	4.01[b]
ATP synthase α	543	3.774	0.025	0.130	0.191	0.427	0.007	0.006	0.006	2.41
ATP synthase β	357	2.568	0.017	0.095	0.165	0.367	0.000	0.005	0.005	1.72
β-1, 4-galactosyl transferase	396	2.552	0.210	0.103	0.248	0.342	0.025	0.068	0.052	1.96
Carboxypeptidase	432	1.461	0.045	0.142	0.251	0.529	0.004	0.020	0.017	2.56
Connexin	381	1.773	0.020	0.236	0.137	0.469	0.006	0.005	0.006	1.13
Corticotropin-releasing factor	182	3.720	0.184	0.182	0.409	0.523	0.018	0.110	0.083	1.37
Dopamine receptor D2	442	3.753	0.027	0.235	0.254	0.385	0.009	0.005	0.009	0.71
Fibrinogen α	433	2.735	0.200	0.114	0.210	0.604	0.033	0.072	0.072	8.00[c]
Glucose transporter	491	5.080	0.023	0.113	0.320	0.562	0.009	0.006	0.008	4.18[b]
Growth hormone	189	4.172	0.126	0.738	0.333	0.465	0.168	0.032	0.029	3.67[b]
Growth hormone receptor	636	2.997	0.398	0.053	0.164	0.344	0.059	0.030	0.133	9.94[c]
Hexokinas I	915	1.915	0.073	0.166	0.226	0.550	0.018	0.029	0.020	14.99[c]
IGF binding protein 1	258	2.605	0.195	0.307	0.460	0.667	0.109	0.082	0.084	3.14[b]
IGF binding protein 3	287	3.463	0.122	0.046	0.755	0.612	0.044	0.055	0.061	6.46[c]
Insulin-like growth factor 1	114	3.136	0.036	0.019	0.325	0.564	0.004	0.008	0.020	0.40
Insulin-like growth factor 2	149	3.385	0.122	0.199	0.431	0.332	0.025	0.049	0.043	0.04
Interleukin 1α	260	2.764	0.467	0.181	0.147	0.374	0.078	0.086	0.161	0.29
Interleukin 1β	263	2.424	0.435	0.131	0.349	0.375	0.082	0.168	0.118	1.31
Interleukin 2	152	3.341	0.665	0.061	0.121	0.646	0.047	0.217	0.226	7.44[c]
Interleukin 6	205	3.057	0.838	0.100	0.255	0.566	0.191	0.178	0.373	1.66
Interleukin 7	153	2.371	0.591	0.067	0.101	0.296	0.098	0.069	0.097	3.43[b]
Lactate dehydrogenase A	331	2.576	0.066	0.115	0.136	0.581	0.020	0.017	0.015	9.08[c]
Lactoferrin	662	2.572	0.313	0.168	0.407	0.481	0.068	0.131	0.127	1.17
Luteinizing hormone receptor	685	3.495	0.201	0.120	0.139	0.376	0.042	0.030	0.052	6.26[c]
Myelin proteolipid protein	148	2.059	0.083	0.033	0.077	0.117	0.009	0.009	0.000	3.93[b]
Neuroleukin	557	1.944	0.083	0.287	0.197	0.460	0.017	0.018	0.045	1.14
Neurophysin I	162	4.666	0.072	0.296	0.258	0.989	0.023	0.025	0.060	0.34
Neurophysin II	116	2.985	0.044	0.131	1.127	0.992	0.037	0.063	0.008	3.33[b]
Ornithine decarboxylase	460	2.544	0.091	0.257	0.222	0.311	0.016	0.018	0.038	1.69
Plasminogen activator inhibitor	386	3.149	0.125	0.234	0.329	0.692	0.033	0.041	0.083	0.08
Prolactin	197	2.550	0.355	0.162	0.389	0.523	0.053	0.109	0.230	0.76
Proopiomelanocortin	211	4.557	0.050	0.313	0.513	0.919	0.014	0.019	0.058	0.50
Protein disulfide isomerase	505	2.364	0.046	0.246	0.395	0.567	0.017	0.012	0.025	1.49
Terminal transferase	506	2.656	0.228	0.150	0.093	0.477	0.042	0.035	0.081	3.82[b]
Thrombomodulin	341	2.990	0.143	0.414	0.567	1.337	0.092	0.112	0.108	4.69[c]
Transforming growth factor β1	315	2.903	0.061	0.304	0.336	0.684	0.014	0.016	0.054	1.02
Transforming growth factor β2	413	3.052	0.033	0.115	0.179	0.408	0.003	0.001	0.019	3.14[b]
Transforming growth factor β3	408	3.597	0.068	0.111	0.259	0.362	0.002	0.041	0.009	12.17[c]
Transforming growth factor β3 receptor	843	2.943	0.162	0.133	0.386	0.421	0.038	0.052	0.060	3.70[b]
Urokinase-plasminogen activator	403	2.108	0.356	0.196	0.208	0.354	0.078	0.066	0.125	0.18
Mean	388	2.934	0.173	0.190	0.291	0.525	0.039	0.051	0.066	3.36

[a] L_c: number of codons in the gene; κ: transition/transversion rate ratio, estimated under model A (Fig. 1); ω; nonsynonymous/synonymous rate ratio, estimated under model A (Fig. 1); d_S and d_N: numbers of synonymous and nonsynonymous substitutions per site, respectively, calculated for the primate (P), artiodactyl (A), and rodent (R) branches of Fig. 1; Δℓ: log-likelihood difference between models A and B (Fig. 1)

[b] Significant at 5% level, $\chi^2 = 3.00$

[c] Significant at 1% level, $\chi^2 = 4.61$

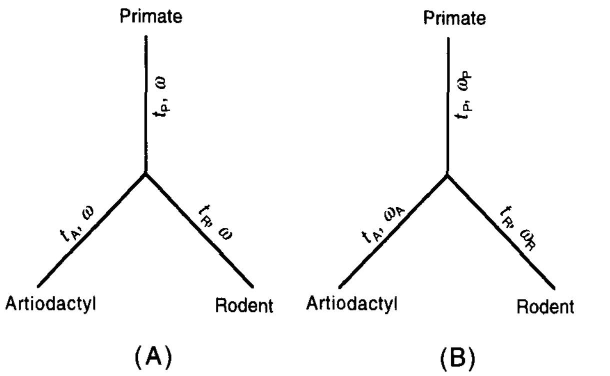

Fig. 1. The unrooted phylogenetic tree of primates, artiodactyls, and rodents showing two evolutionary models. The *t*s are branch lengths, while ω ($= d_N/d_S$) is the nonsynonymous/synonymous rate ratio. In **A**, the d_N/d_S ratio is constrained to be constant among branches (lineages) in the tree. In **B**, the ratios are allowed to vary among branches, and a different ω parameter is used for each branch.

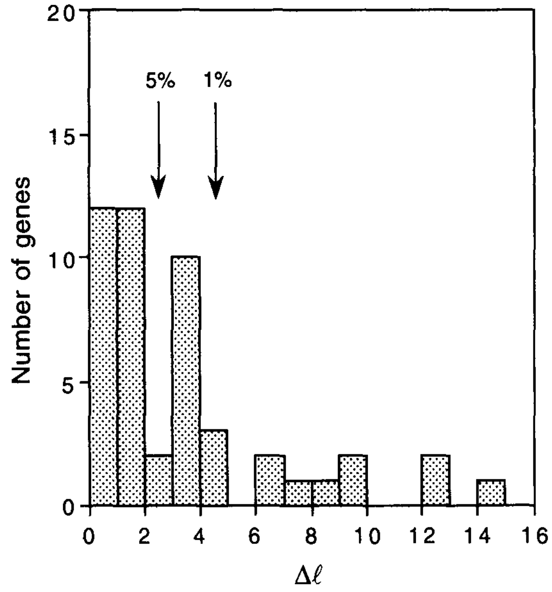

Fig. 2. The frequency distribution, among 48 nuclear genes, of the log-likelihood difference between models **A** and **B** of Fig. 1. The *arrows* designate the critical values at the 5% and 1% significance levels.

Estimated numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site for the three branches in the tree of Fig. 1 are obtained under model B (Table 2). In growth hormone receptor and interleukins 6 and 7, the estimated d_N/d_S ratios for the primate lineage are greater than one and are 1.11, 1.91, and 1.46, respectively. These genes may be under positive Darwinian selection (Ohta 1995). The averages of estimates of d_S among genes are 0.190, 0.291, 0.525 for the primate, artiodactyl, and rodent branches, respectively (Table 2); these values are in the proportion 0.57:0.87:1.57, not very different from the weighting factors calculated by Ohta (1995: Table 2), that is, 0.61:0.82:1.58. The averages of estimates of d_N among genes are 0.039, 0.051, and 0.066 for the primate, artiodactyl, and rodent branches, respectively (Table 2); these are in the proportion 0.75:0.98:1.27, almost identical to the weighting factors calculated by Ohta, that is, 0.75:0.97:1.28. The synonymous rates are more variable among lineages than the nonsynonymous rates, which conforms with the previous observation that the lineage effect is more pronounced for synonymous substitutions than for nonsynonymous substitutions (Gillespie 1991; Ohta 1993, 1995). The ratios of the average synonymous and nonsynonymous rates ($\bar{d}_N/\bar{d}_S$) are 0.21, 0.17, and 0.13 for the primate, artiodactyl, and rodent lineages, respectively, while the average of the ratios ($\bar{\omega}$) are 0.28, 0.22, and 0.13, respectively. These estimates are in general agreement with the observation of Ohta (1993) that the primate lineage has, on average, reduced synonymous and nonsynonymous rates and appears to be under less severe purifying selection, possibly due to reduced population sizes. Nevertheless, the average patterns over genes are not very illuminating as substantial variation exists among genes (Table 2).

The difference of log-likelihood values under models A and B (Fig. 1), $\Delta\ell$, is calculated for each gene and listed in Table 2, and the distribution of $\Delta\ell$ among the 48 genes is shown in Fig. 2. If the d_N/d_S ratio (ω) is constant among lineages, that is, if model A is correct, $2\Delta\ell$ should follow a χ^2 distribution with *d.f.* = 2. By this likelihood ratio test, the d_N/d_S ratio is significantly variable among lineages in 12 genes ($P < 5\%$) and is highly significantly variable in ten other genes ($P < 1\%$). In sum, about half of the genes show significantly different d_N/d_S ratios among lineages, and the results provide strong evidence for rejecting the strictly neutral model.

The same analysis was performed using the 61 codon frequencies as free parameters. This model gives significantly better fit to data of every gene than using nucleotide frequencies at the codon positions, judged by a χ^2 critical value with *d.f.* = 60 − 9 = 51. However, the likelihood difference between models A and B (Fig. 1) changed very little with this change of assumptions about codon frequencies, and the likelihood ratio test of constancy of d_N/d_S among lineages was found to be quite robust to this aspect of the model assumptions. We presented results obtained from the simpler model because the estimates are similar and because some genes are short and may not provide enough information to estimate many parameters.

Although our general conclusions agree with those of Ohta (1995), considerable differences exist between our likelihood estimates of the synonymous substitution rates (Table 2) and the corresponding estimates of Ohta (1995: Table 3), who used Ina's (1995) method. For example, for the first gene (acetylcholine receptor α), our estimates of d_S are 0.147, 0.165, and 0.404 for the primate, artiodactyl, and rodent lineages, respectively, while Ohta's estimates are 0.161, 0.161, and 0.296 (misprinted as 0.116, 0.116, and 0.296). The nonsynonymous rate estimates are much more similar, and are 0.013, 0.007, 0.018 for the three branches, respectively, by our likelihood analysis, and are 0.014, 0.007, 0.019 by Ohta's

Table 3. Pairwise comparisons of the acetylcholine receptor α genes of primates, artiodactyls and rodents[a]

Model		κ	S	N	d_S	d_N	d_N/d_S (ω)
Primate–rodent							
1	ML,Fequal	1	349.3	1,021.7	0.496	0.029	0.059
2	ML,Fequal	2.78	397.6	973.4	0.421	0.031	0.073
3	ML,F1 × 4	2.89	407.2	963.8	0.436	0.031	0.071
4	ML,F3 × 4	3.02	328.8	1,042.2	0.546	0.030	0.055
5	ML,F61	2.99	318.5	1,052.5	0.615	0.030	0.048
6	NG		321.2	1,049.8	0.523	0.030	0.058
7	Ina	6.08	408.4	962.6	0.405	0.033	0.081
Primate–artiodactyl							
1	ML,Fequal	1	349.3	1,021.7	0.288	0.018	0.064
2	ML,Fequal	3.16	403.5	967.5	0.241	0.020	0.081
3	ML,F1 × 4	3.33	413.8	957.2	0.253	0.019	0.077
4	ML,F3 × 4	3.44	320.8	1,050.2	0.312	0.019	0.060
5	ML,F61	3.35	306.1	1,064.9	0.341	0.018	0.054
6	NG		320.9	1,050.1	0.294	0.019	0.066
7	Ina	6.39	409.5	961.5	0.228	0.021	0.092
Artiodactyl–rodent							
1	ML,Fequal	1	349.3	1,021.7	0.488	0.023	0.047
2	ML,Fequal	2.44	391.4	979.6	0.425	0.024	0.057
3	ML,F1 × 4	2.60	403.8	967.2	0.444	0.024	0.055
4	ML,F3 × 4	2.67	316.1	1,054.9	0.564	0.024	0.042
5	ML,F61	2.61	304.9	1,066.1	0.636	0.023	0.036
6	NG		321.4	1,049.6	0.520	0.024	0.046
7	Ina	5.28	402.7	968.3	0.404	0.026	0.064

[a] In the maximum likelihood (ML) analysis, codon frequencies are assumed to be equal (Fequal), or calculated from nucleotide frequencies (F1 × 4), nucleotide frequencies at codon positions (F3 × 4), or treated as free parameters (F61). NG: method of Nei and Gojobori (1986). Ina: method of Ina (1995)

analysis. An extreme case is the thrombomodulin gene, for which our likelihood estimates of d_S for the three branches are 0.414, 0.567, 1.337 (Table 2), while Ohta's estimates are 0.169, 0.237, 0.570; the estimates from the two analyses differ by a factor greater than two. Estimates of d_N for the three branches are 0.092, 0.112, 0.108 by the likelihood method (Table 2) and 0.112, 0.125, 0.125 by Ohta's analysis. Generally, our likelihood estimates of d_S are greater than Ohta's while our estimates of d_N are very similar to, but smaller than, Ohta's.

The two analyses differ in many ways. One difference is that in the likelihood analysis, gene sequences from all three species are compared simultaneously, while Ina's (1995) method calculates the numbers of synonymous and nonsynonymous substitutions per site (d_S and d_N) for each pairwise comparison, which are then used to estimate rates for branches. Let the pairwise distances be d_{PA}, d_{PR}, d_{AR}, where d can be either d_S or d_N. Ohta calculated the numbers of substitutions for the three branches in the tree (Fig. 1) as

$$\begin{aligned} d_P &= (d_{PA} + d_{PR} - d_{AR})/2 \\ d_A &= (d_{PA} + d_{AR} - d_{PR})/2 \qquad (4) \\ d_R &= (d_{PR} + d_{AR} - d_{PA})/2 \end{aligned}$$

While this calculation may introduce some bias into the estimates when the d_N/d_S ratios vary among lineages, it does not seem to be the major reason for the differences between our likelihood estimates and Ohta's results. The differences between the methods are explored in the next section.

Estimation of Synonymous and Nonsynonymous Substitution Rates

In order to understand the differences in estimates of d_S and d_N between our likelihood analysis and Ina's (1995) method, which Ohta (1995) used, we analyze several genes (acetylcholine receptor α, lactate dehydrogenase A, and thrombomodulin) in more detail. The method of Nei and Gojobori (1986) (NG) is also used for comparison. The NG method and Ina's method were designed for estimating d_S and d_N between two sequences. Therefore, we also apply the likelihood method to the three pairwise comparisons for each gene, varying assumptions about the transition/transversion rate ratio (κ) and the codon frequencies. The patterns obtained for different genes are similar, and results for two genes only (acetylcholine receptor α and thrombomodulin) are presented (Tables 3 and 4). Sequence divergence between the acetylcholine receptor α genes is lower than average, but the thrombomodulin genes are very divergent among species and also have high d_N/d_S ratios (Table 2).

Several different models were assumed in the likelihood analysis concerning the codon frequency parameters (π_j in Eq. 1). The most parameter-rich version (F61) uses each codon frequency as a parameter, with the

Table 4. Estimation of the numbers of synonymous and nonsynonymous substitutions per site in pairwise comparison of the thrombomodulin genes of primates, artiodactyls, and rodents[a]

Model		κ	S	N	d_S	d_N	d_N/d_S (ω)
Primate–rodent							
1	ML,Fequal	1	260.6	762.4	0.997	0.199	0.200
2	ML,Fequal	2.14	287.3	735.7	0.864	0.209	0.241
3	ML,F1 × 4	2.59	318.2	704.8	1.111	0.208	0.188
4	ML,F3 × 4	2.69	255.8	797.2	0.658	0.200	0.120
5	ML,F61	3.01	252.5	770.5	2.307	0.216	0.094
6	NG		239.4	783.6	0.976	0.211	0.216
7	Ina	4.19	298.5	724.5	0.741	0.232	0.313
Primate–artiodactyl							
1	ML,Fequal	1	260.6	762.4	0.517	0.215	0.417
2	ML,Fequal	2.18	288.0	735.0	0.453	0.226	0.500
3	ML,F1 × 4	2.75	326.3	696.7	0.652	0.219	0.336
4	ML,F3 × 4	3.08	156.9	866.1	1.231	0.204	0.166
5	ML,F61	3.43	165.7	857.3	1.332	0.219	0.164
6	NG		243.3	779.7	0.511	0.219	0.428
7	Ina	3.82	297.8	725.2	0.411	0.238	0.580
Artiodactyl–rodent							
1	ML,Fequal	1	260.6	762.4	1.098	0.217	0.198
2	ML,Fequal	2.16	287.6	735.4	0.942	0.228	0.242
3	ML,F1 × 4	2.59	318.0	705.0	1.211	0.229	0.189
4	ML,F3 × 4	2.71	231.9	791.1	1.938	0.221	0.114
5	ML,F61	3.12	256.5	766.5	2.697	0.242	0.090
6	NG		240.9	782.1	1.091	0.228	0.209
7	Ina	4.68	303.7	719.3	0.815	0.252	0.309

[a] See note to Table 3

only constraint that the sum is one; with the standard genetic code, 61 − 1 = 60 free parameters are needed. A less parameter-rich version uses different base frequencies at the three codon positions (F3 × 4) to calculate the expected codon frequencies, and needs 3 × (3 − 1) = 9 free parameters. If the differences in base frequency distribution at the three codon positions are ignored, the four nucleotide frequencies (F1 × 4) can be used to calculate the expected codon frequencies, with only three free parameters needed. In all these models, the frequency parameters are estimated using the observed codon or base frequencies. The simplest case is to assume equal frequency for each codon (Fequal). The Fequal, F1 × 4, F3 × 4, and F61 models are nested and can be compared using likelihood ratio tests. The results (not shown) of such tests suggest that for each gene, the simpler models are rejected in favor of their more complex alternatives. Codon frequencies are unequal and cannot be predicted from nucleotide frequencies. The different models are used to examine their effects on estimation of d_S and d_N. In the following, the likelihood analyses under different model assumptions are regarded as different methods.

We first examine the performance of the NG method, which is an approximate implementation of the Fequal model without transition/transversion rate bias. The results obtained using the NG method are indeed very similar to the likelihood results under that model (Fequal, κ = 1) (Tables 3 and 4). The differences in estimates of S and N between the two methods are due to the fact that the expected codon frequencies (π_j = 1/61 for all j) are used in the likelihood calculation, while the NG method counts the number of sites (S and N) codon by codon and effectively uses the observed codon frequencies. If the same codon frequencies are used in both methods, the two methods produce identical estimates of S and N. For all three genes examined, the estimates of S by the likelihood analysis are approximately 7–8% greater than the estimates by the NG method (Tables 3 and 4). The proportion of synonymous sites in the gene is not as high as expected under the assumption of equal codon frequencies. Apparently, amino acids coded by four codons do not occur twice as often as amino acids coded by two codons (results not shown).

Compared with the likelihood estimates, the NG method tends to underestimate the number of synonymous substitutions for the entire sequence (Sd_S) and overestimate the number of nonsynonymous substitutions (Nd_N). For example, for the primate–rodent comparison of the acetylcholine receptor α gene, the estimates of Sd_S and Nd_N are 168.0 and 31.5, respectively, by the NG method and 173.2 and 29.6 by likelihood (Table 3), with 3–6% differences between methods. The major reason for this difference appears to be the equal weighting of substitution pathways used in the NG method to count the numbers of differences between codons differing at two or three positions. When $d_N/d_S <$ 1, pathways involving synonymous changes are more likely than those involving nonsynonymous changes, and equal weighting tends to underestimate S_d and overestimate N_d, as pointed out by Miyata and Yasunaga (1980) and Li et al. (1985). Appropriate weighting of pathways requires knowledge of the d_N/d_S ratio, which is being estimated. In the above comparison, 14 out of 456 codons are different at two positions in the two genes. This problem becomes more serious when more divergent genes are compared. For example, in the comparison of the primate and rodent thrombomodulin genes, 54 out of 341 codons are different at two positions, and 12 codons are different at all three positions in the two species. Estimates of Sd_S and Nd_N between the two sequences are 233.7 and 165.3, respectively, by the NG method and 259.8 and 151.7 by likelihood. The two methods differ by about 10%. When $d_N/d_S > 1$, use of equal weighting is expected to overestimate S_d and underestimate N_d.

Ina's (1995) method is an improvement over the NG method and accounts for the transition/transversion bias. It is well known that ignoring the transition/transversion bias causes underestimation of the number of synonymous sites (S), overestimation of the synonymous rate (d_S), and underestimation of the d_N/d_S ratio (Li 1993; Pamilo and Bianchi 1993). This pattern is apparent when estimates from the two methods are compared (Tables 3 and 4). Results obtained using Ina's method are also very similar to those obtained from the likelihood method assuming equal codon frequencies, with the transition/transversion rate bias estimated (Fequal). Estimates of d_N/d_S by the two methods are the largest among all estimates (Tables 3 and 4). Like the NG method, Ina's method counts the numbers of differences between codons using equal weights for synonymous and nonsynonymous substitutions and tends to overestimate the numbers of nonsynonymous differences and substitutions (N_d and Nd_N). For example, in the primate–rodent comparison of the thrombomodulin genes, the estimates of the total numbers of synonymous and nonsynonymous substitutions (Sd_S and Nd_N) are 221.2 and 168.1, respectively, by Ina's method, while the likelihood estimates are 248.2 and 153.8 (Table 4). The two methods are about 10% different.

Ina's method also uses the observed codon frequencies to count the number of sites (S and N), and may be expected to give smaller estimates of S than the corresponding likelihood model (Fequal, with κ estimated). This is not the case, however. The reason is that a biased estimate of κ is used in Ina's method. Ina (1995) used the third codon positions to estimate the "mutational" transition/transversion rate ratio (κ in Eq. 1), applying the correction of Kimura (1980). The assumption is that mutations at the third position are synonymous. However, as noted by Ina, transitions at the third position are more likely to be synonymous than transversions, and purifying selection will have elevated the transition/transversion rate ratio at the third codon position. For example, for the primate–rodent comparison of the acetylcholine receptor α gene, the likelihood estimate of κ is 2.8, while the estimate obtained by Ina's method is 6.1 (Table 3). The proportion of synonymous sites, $S/(S + N)$, is very sensitive to the assumed value of κ, especially when κ is small (Fig. 3), and use of a large κ leads to overestimation of S, underestimation of d_S, and overestimation of the d_N/d_S ratio. The effect of the overestimated κ is considerable in Ina's method but is counteracted by the use of the observed codon frequencies, so the differences shown in Tables 3 and 4 between Ina's method and the likelihood analysis are not so great. Ina's method has thus overcorrected the NG method by using a biased estimate of the transition/transversion rate ratio. This result seems to explain the perplexing results in Ina's (1995: Table 15) simulations—that Ina's method performed well with a high transition/transversion mutation bias, that is, under the influenza virus and mitochondrial mutation schemes where κ was 3.9 or 5.6 but overestimated by Ina's method as 6.4 or 7.8, and performed more poorly with a low transition/transversion bias, that is, under the nuclear pseudogene mutation scheme where κ was 1.2 but overestimated as 2.0. The reason appears to be that the $S/(S + N)$ ratio is much more sensitive to κ when κ is small than when κ is large (Fig. 3).

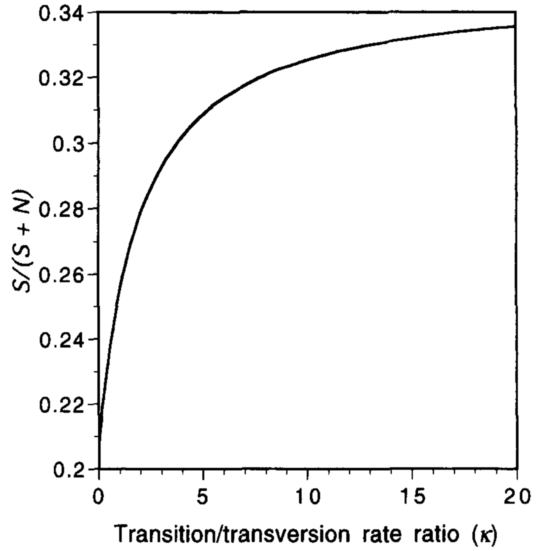

Fig. 3. The proportion of synonymous sites as a function of the transition/transversion mutation rate ratio (κ) when all 61 sense codons in the standard genetic code are equally frequent. The method of Goldman and Yang (1994) is used. The method of Ina (1995) will give exactly the same results if changes (mutations) to stop codons are properly disallowed (see Discussions).

Another complexity of the approximate methods concerns counting synonymous and nonsynonymous sites in codons that can mutate into stop codons in one step. We suggest that the correct approach is to disallow mutations to stop codons and lose some sites (mutational potentials) so that $S + N < 3L_c$. For example, by a one-step mutation at the third position, codon TGT (cytosine) can change into TGC (cytosine), TGA (stop), or TGG (tryptophan). Without transition/transversion bias (the underlying model of the NG method), the third position of the codon should be counted as one-third synonymous sites and one-third nonsynonymous sites, with one-third sites lost. Disallowing mutations to stop codons in this way leads to approximately 4.2% loss of sites under the standard genetic code without transition/transversion bias. Ina (1995: Table 1) counted the third position of TGT as half synonymous sites and half nonsynonymous sites; this effectively raises the mutation rate at that codon position and is not justified. To calculate results in Tables 3 and 4, we have scaled S and N so that they sum to $3L_c$ for all methods except that of Ina; d_S and d_N are thus underestimated by about 4.2% for these methods, although the d_N/d_N ratios are correctly estimated.

Discussion

Comparisons of the likelihood estimates of d_S and d_N with those obtained from the approximate methods of Nei and Gojobori (1986) and Ina (1995) suggest that the NG method is more or less a reliable approximation to the likelihood method assuming equal codon frequencies and ignoring transition/transversion rate bias. Ina's (1995) method is a good approximation to the likelihood method assuming equal codon frequencies and accounting for the transition/transversion rate bias. The same comparisons, however, also highlight some problems of the approximate methods, regarding both the counting of sites (S and N) and the counting of differences (S_d and N_d). The NG method ignores the transition/transversion rate bias and underestimates S, while Ina's method uses an overestimated transition/transversion rate ratio and overestimates S, and has thus overcorrected the NG method. Ina (1995) discussed this problem in great detail and indeed suggested an iterative algorithm to estimate the transition/transversion rate ratio (method 2). Unfortunately, method 2 did not perform consistently better than method 1 in Ina's extensive simulations. In counting the numbers of differences (S_d and N_d), both methods apply equal weights to pathways involving synonymous and nonsynonymous changes and tend to overestimate the number of nonsynonymous differences when $d_N/d_S <$ 1. This problem exists when the level of sequence divergence is not very low and some codons differ in the two sequences at more than one codon position.

Furthermore, estimates of d_S and d_N obtained from the likelihood analysis under different model assumptions may vary considerably (Tables 3 and 4). This unfortunate result suggests that the estimation of d_S and d_N by both the likelihood methods and the approximate methods is sensitive to assumptions about the transition/transversion rate bias and codon frequency bias. We note that, based on his simulation results, Ina (1995) also cautioned on the use of the approximate methods in the presence of extreme base or codon frequency biases. The codon frequency bias usually has different effects from the transition/transversion rate bias, which leads to the ironic result that the NG method gives estimates that are closer to the likelihood results under more-realistic models (such as F3 × 4 and F61) than does Ina's method (Tables 3 and 4). While this is the case for all the pairwise comparisons of the three genes examined, the general pattern concerning the reliability of the two methods remains uncertain. In the absence of approximate methods that can properly take into account factors such as the transition/transversion rate bias and codon frequency bias, it is advisable to use the likelihood method, even for pairwise sequence comparison, as more realistic assumptions can be easily implemented in the likelihood analysis. It may also be argued that the likelihood approach to estimating d_S and d_N is conceptually simpler than the approximate methods, as it does not involve the ad hoc treatments of the approximate methods in counting the numbers of differences. Correction for multiple hits is also taken care of automatically by the maximum likelihood methodology, and d_S and d_N are calculated directly from their definitions. A disadvantage of the likelihood method is its computational requirement, which is about 30 s for a pairwise comparison on a PowerMac 8500/120, while calculation by the approximate methods is completed almost instantaneously.

The d_N/d_S ratio is found to vary significantly among lineages in 22 of the 48 gene loci examined. This result provides strong evidence against a strictly neutral model of molecular evolution, i.e., a model involving strictly neutral and deleterious mutations only. Analyses similar to Ohta (1995) using the dispersion index could be performed with the estimated synonymous and nonsynonymous rates of Table 2. This is not pursued here as there are only three lineages in the data, and calculation of the variance of the numbers of substitutions among lineages and the dispersion index may not be very reliable. Nevertheless, we note that our results confirm the claim of Ohta (1995) and Gillespie (1991) that a strictly neutral model is not an adequate description of the evolutionary processes of the genes examined. Despite inaccuracies of the analytical procedures used by Gillespie and Ohta, their major conclusions regarding the mechanisms of molecular evolution appear to be justified.

Acknowledgments. We thank T. Ohta for sending the sequence data analyzed in this paper, and W.-H. Li and Y. Ina for comments. This study is supported by NIH grant GM40282 to Montgomery Slatkin and a personal grant to R.N. from the Danish Research Council.

References

Cameron JM (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol 41: 1152–1159

Easteal S, Collet C (1994) Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: protein evolution in mammals is not neutral. Mol Biol Evol 11:643–647

Eyre-Walker A, Gaut BS (1997) Correlated rates of synonymous site evolution across plant genomes. Mol Biol Evol 14:455–460

Gillespie JH (1987) Molecular evolution and the neutral allele theory. Oxf Surv Evol Biol 4:10–37

Gillespie JH (1989) Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol 6:636–647

Gillespie JH (1991) The causes of molecular evolution. Oxford University Press, Oxford

Goldman N (1994) Variance to mean ratio, *R(t)*, for Poisson processes on phylogenetic trees. Mol Phylogenet Evol 3:230–239

Goldman N, Yang Z (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. Science 185:862–864

Ina Y (1995) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol 40:190–226

Ina Y (1996) Patterns of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. J Genet 75:91–115

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

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

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila.* Nature 351:652–654

Miyata T, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. J Mol Evol 16:23–36

Muse SV, Gaut BS (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, Gojobori T (1986) Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Nielsen R (1997) Robustness of the estimator of the index of dispersion for DNA sequences. Mol Phylogenet Evol 7:346–351

Ohta T (1993) A examination of the generation-time effect on molecular evolution. Proc Natl Acad Sci USA 90:10676–10680

Ohta T (1995) Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol 40:56–63

Pamilo P, Bianchi NO (1993) Evolution of the *Zfx* and *Zfy* genes—rates and interdependence between the genes. Mol Biol Evol 10: 271–281

Yang Z (1997) Phylogenetic analysis by maximum likelihood (PAML), version 1.3. University of California, Berkeley, California, USA