J Mol Evol (1999) 48:274–283

JOURNAL OF MOLECULAR EVOLUTION

Estimation of the Transition/Transversion Rate Bias and Species Sampling

Ziheng Yang,[1] Anne D. Yoder[2,3]

[1]Department of Biology (Galton Laboratory), University College London, 4 Stephenson Way, London NW1 2HE, England
[2]Department of Cell and Molecular Biology, Northwestern University Medical School, Chicago, IL 60611, USA
[3]Department of Zoology, Field Museum of Natural History, Chicago, IL 60605, USA

Received: 1 October 1997 / Accepted: 29 September 1998

Abstract. The transition/transversion (ti/tv) rate ratios are estimated by pairwise sequence comparison and joint likelihood analysis using mitochondrial cytochrome *b* genes of 28 primate species, representing both the Strepsirrhini (lemurs and lories) and the Anthropoidea (monkeys, apes, and humans). Pairwise comparison reveals a strong negative correlation between estimates of the ti/tv ratio and the sequence distance, even when both are corrected for multiple substitutions. The maximum-likelihood estimate of the ti/tv ratio changes with the species included in the analysis. The ti/tv bias within the lemuriform taxa is found to be as strong as in the anthropoids, in contradiction to an earlier study which sampled only one lemuriform. Simulations show the surprising result that both the pairwise correction method and the joint likelihood analysis tend to overcorrect for multiple substitutions and overestimate the ti/tv ratio, especially at low sequence divergence. The bias, however, is not large enough to account for the observed patterns. Nucleotide frequency biases, variation of substitution rates among sites, and different evolutionary dynamics at the three codon positions can be ruled out as possible causes. The likelihood-ratio test suggests that the ti/tv rate ratios may be variable among evolutionary lineages. Without any biological evidence for such a variation, however, we are left with no plausible explanations for the observed patterns other than a possible saturation effect due to the unrealistic nature of the model assumed.

Correspondence to: Ziheng Yang; *e-mail:* z.yang@ucl.ac.uk

Key words: Transition/transversion rate ratio — Maximum likelihood — Species sampling — Saturation — Lemurs — Primates

Introduction

In virtually all DNA sequences from any genome examined, transitions (T↔C, A↔G) have been noted to occur at higher frequencies than transversions (T↔A, T↔G, C↔A, C↔G) (Brown et al. 1982; Gojobori et al. 1982; Curtis and Clegg 1984; Wakeley 1994, 1996). While transition/transversion (ti/tv) bias is known to be a general property of DNA sequence evolution, it is more pronounced in animal mitochondrial DNAs (mtDNAs) than in nuclear or chloroplast DNAs (for a review see Wakeley 1996). Estimation of the ti/tv rate bias is important not only to our understanding of the patterns of DNA sequence evolution, but also to reliable estimation of sequence distance and phylogeny reconstruction.

The ti/tv rate ratio has most often been estimated by counting the proportions of sites with transitional (P) and transversional (Q) differences between two sequences, that is, by P/Q. A major problem with this approach has been the correlation between the genetic distance and the observed ti/tv (P/Q). It has been repeatedly noted that at low levels of genetic divergence, ti/tv appears to be high, and at high levels of genetic divergence, ti/tv appears to be low. Figure 1 illustrates the pattern in three eutherian groups. In all three, transition frequencies (P) appear to peak at approximately 15% genetic divergence, whereupon they are gradually matched in frequency by transversions as genetic distance increases. At levels of 20%

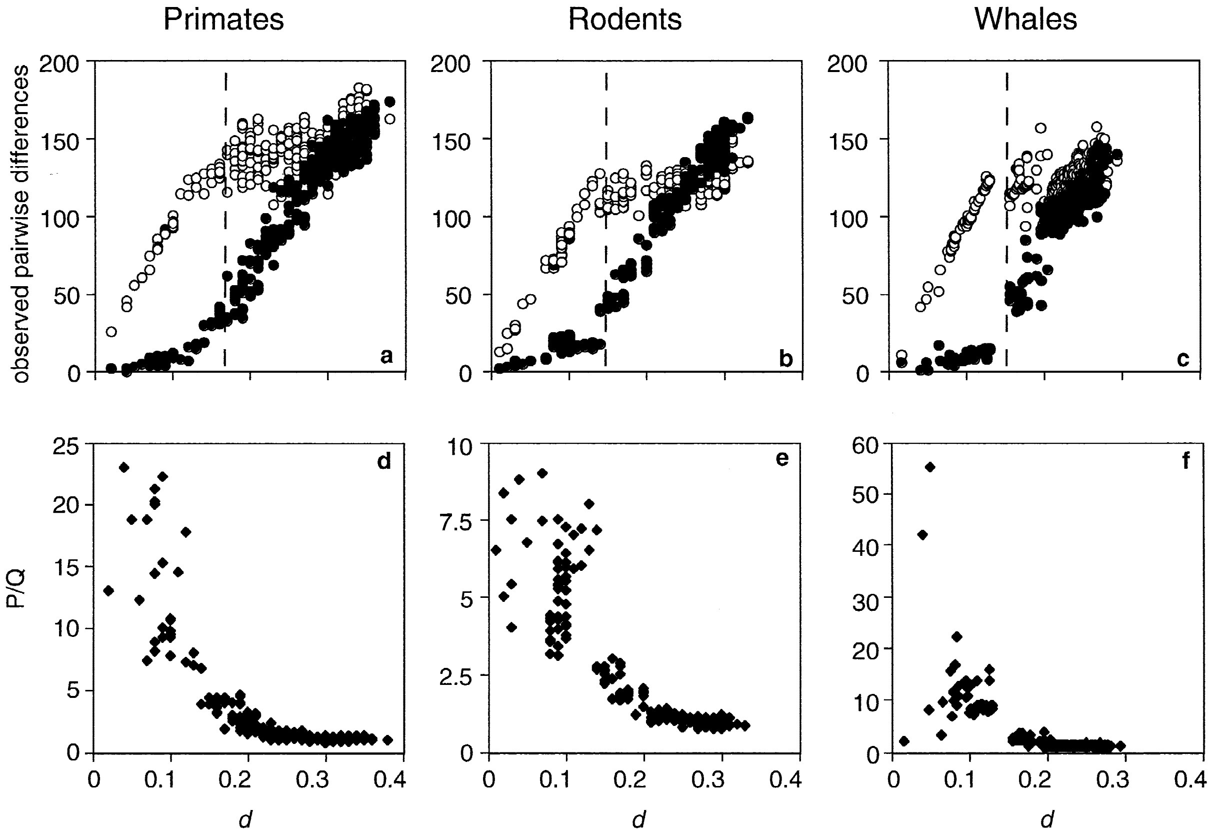

Fig. 1. Negative correlation between apparent (uncorrected) ti/tv and genetic divergence in cytochrome *b* from three mammalian groups. **A–C** Counts of sites with transitional (*open circles*) and transversional (*filled circles*) differences in pairwise comparisons relative to sequence divergence corrected with the K80 model. *Dashed line* indicates the level of sequence divergence at which transitions appear to approach saturation. **D–F** *P/Q,* derived from uncorrected pairwise transition and transversion frequencies. The three data sets are used to show the generality of the problem in using *P/Q* to estimate ti/tv, but only the primate data set is analyzed in detail in this paper. Primate cytochrome *b* sequences used in A and D are specified in Table 1. Species and GenBank accession numbers (in parentheses) for rodent sequences used in B and E are *Acomys cahirinus* (Z96051), *A. cahirinus* (Z96052, Z96053), *A. dimidiatus* (Z96060, Z96061, Z96062), *A. ignitus* (Z96063, Z96064), *A. russatus* (Z96065, Z96066), *A. sp.* (Z96055, Z96056, Z96057, Z96058, Z96059), *A. spinosissimus* (Z96068), *Mus musculoides* (Z96069), *Mus musculus* (J01420), *Ctenomys sociabilis* (U34853), *Euryzygomatomys spinosus* (U34858), *Spermophilus richardsonii* (S73150), *Thrichomys apereoides* (U34854), *Trinomys albispinis* (U34856), and *Rattus norvegicus* (X14848). In Figs. C and F cetaceans with artiodactyl and perissodactyl outgroups are analyzed. The species names and GenBank accession numbers (in parentheses) are *Balaenoptera musculus* (X72204), *B. acutorostrata* (X75753), *B. bonaerensis* (X75581), *B. borealis* (X75582), *B. edeni* (X75583), *B. glacialis* (X75887), *Balaena mysticetus* (X75588), *Caperea marginata* (X75586), *Eschrichtius robustus* (X75585), *Megaptera novaeangliae* (X75584), *Physeter macrocephalus* (X75589), *Stenella attenuata* (X56294), *S. longirostrus* (X56292), *Halichoerus grypus* (X72004), *Dugong dugong* (U07564), *Hippopotamus amphibius* (U07565), *H. amphibius* (Y08813), *H. liberiensis* (Y08814), *Sus scrofa* (X56295), *Bos taurus* (J01394), *Camellus dromedarius* (X56281), *Antilocapra americana* (X56286), *Equus grevyi* (X56282), *E. caballus* (X79547), and *Diceros bicornis* (X56283).

divergence or more, the two substitution types show equal frequencies, thereby yielding an apparent ti/tv of approximately 1. Saturation of transitions at high levels of genetic divergence is commonly believed to explain the pattern (e.g., Brown et al. 1982; Moritz et al. 1987). Thus, it has been suggested that the comparison of closely related sequences may yield more accurate estimates of substitution patterns than will comparisons of more divergent sequences (e.g., Purvis and Bromham 1997). Unfortunately, there are other confounding factors that limit the power of such comparisons. For example, there will typically be a huge variance in observed ti/tv at low levels of divergence when a limited number of sites are examined for sequences with a high transition bias (Wakeley 1996). Also, there is recent evidence to suggest that different sites can exhibit vastly different ti/tv rate ratios (e.g., Wills 1995). Given these caveats, methods that correct for multiple substitutions should offer the best option for the accurate determination of the ti/tv rate ratio (Jukes 1987).

A maximum-likelihood approach was used by Hasegawa et al. (1990) to estimate the rates of transitions and transversions in primates in an 896-bp segment of mtDNA that contains parts of two proteins (ND4 and ND5) and three tRNA genes (Brown et al. 1982; Hayasaka et al. 1988). The species sample included the hominoids (human, chimpanzee, gorilla, orangutan, gibbon), four closely related cercopithecoids (Japanese macaque, rhesus macaque, crab-eating macaque, and Barbary macaque), one platyrrhine (common squirrel

Table 1. Sources of primate cytochrome *b* sequences analyzed in this paper

Classification	Binomial	GenBank Accession No.	Reference
Lemuriformes			
Lemuridae	*Lemur catta*	U53575	Yoder et al. (1996a)
	Hapalemur griseus	U53574	"
	Eulemur fulvus collaris	U53576	"
	Eulemur fulvus rufus	U53577	"
	Eulemur fulvus albifrons	AF081048	Yoder and Irwin (in press)
	Eulemur macaco macaco	AF081049	"
	Eulemur macaco flavifrons	AF081050	"
	Eulemur mongoz	AF081051	"
	Eulemur rubriventer	AF081052	"
	Varecia variegata rubra	U53578	Yoder et al. (1996a)
Cheirogaleidae	*Cheirogaleus major*	U53570	"
	Mirza coquereli	U53571	"
	Microcebus murinus	U53572	"
Indridae	*Propithecus tattersalli*	U53573	"
Daubentonidae	*Daubentonia madagascariensis*	U53569	"
Lorisiformes			
Galagonidae	*Galago crassicaudatus*	U53579	"
Loridae	*Loris tardigradus*	U53581	"
	Nycticebus coucang	U53580	"
Anthropoidea			
Cebidae	*Saimiri sciureus*	U53582	"
Cercopithecoidea	*Colobus guereza*	U38264	Collura and Stewart (1995)
	Macaca mulatta	U38272	"
Hominoidea	*Hylobates agilis*	U38263	"
	Pongo pygmaeus	U38274	"
	Pongo pygmaeus	D38115	Horai et al. (1995)
	Pan paniscus	D38116	"
	Pan troglodytes	D38113	"
	Gorilla gorilla	D38114	"
	Homo sapiens	J01415	Anderson et al. (1981)

monkey), *Tarsius* (the Philipine tarsier), and one strepsirrhine (the ring-tailed lemur, *Lemur catta*). Hasegawa et al.'s maximum-likelihood analysis allowed different transition rates and transversion rates in different lineages of the phylogeny and found that the transition rate is an order of magnitude lower in the lemur than in the hominoids. The fact that the ring-tailed lemur was the only representative of the Strepsirrhini to be included in the study raises a concern: perhaps the extremely low estimated transition rate is related to the large genetic divergence that separates *Lemur* from the other primates included in the study. Our study addresses this concern by including mitochondrial sequences from numerous closely related lemuriforms. When these sequences are compared, a contradictory pattern emerges. The ti/tv rate bias in the lemuriforms appears to be as pronounced as in the hominoids.

In Hasegawa et al.'s study, the numbers of transitional and transversional differences between pairs of sequences, instead of the original sequence data, were analyzed using a normal approximation. This approximation appears quite good and is unlikely to account for the conflict with our results. Moreover, Hasegawa et al.'s results are reproduced by our study of a different mitochondrial gene, cytochrome *b,* when a similar taxon sample is examined. The purpose of our paper is to describe the surprising patterns in estimates of the ti/tv rate ratios and explore possible causes for the observed patterns. Although we can rule out several molecular evolutionary factors as the proximal cause of the patterns, a biologically plausible interpretation that is consistent with the results remains elusive. Clearly more work is necessary to fully understand the ti/tv rate bias.

Data and Methods

Sequence Data

Mitochondrial cytochrome *b* genes of 28 primate species were examined. The species are identified in Table 1, and one of the most likely phylogenies of the species is shown in Fig. 2. The structure of the strepsirrhine portion of the phylogenetic tree is well corroborated by analyses of both molecular and morphological data (Yoder et al. 1996a, b, 1997), allowing a detailed analysis of the effects of species sampling on the estimation of the ti/tv rate ratio. The position of *Propithecus tattersalli* (Tattersall's sifaka) is locally uncertain and trees in which this species clusters with either of node *a* (as in the tree in Fig. 2), *b*, or *c* are also likely to be true. The three trees produced almost-identical results for the analyses discussed in this paper, and only those using the tree in Fig. 2 are presented. The sequence alignment contains 1140 nucleotide sites, although 7 sites involving undetermined nucleotides are removed from all sequences.

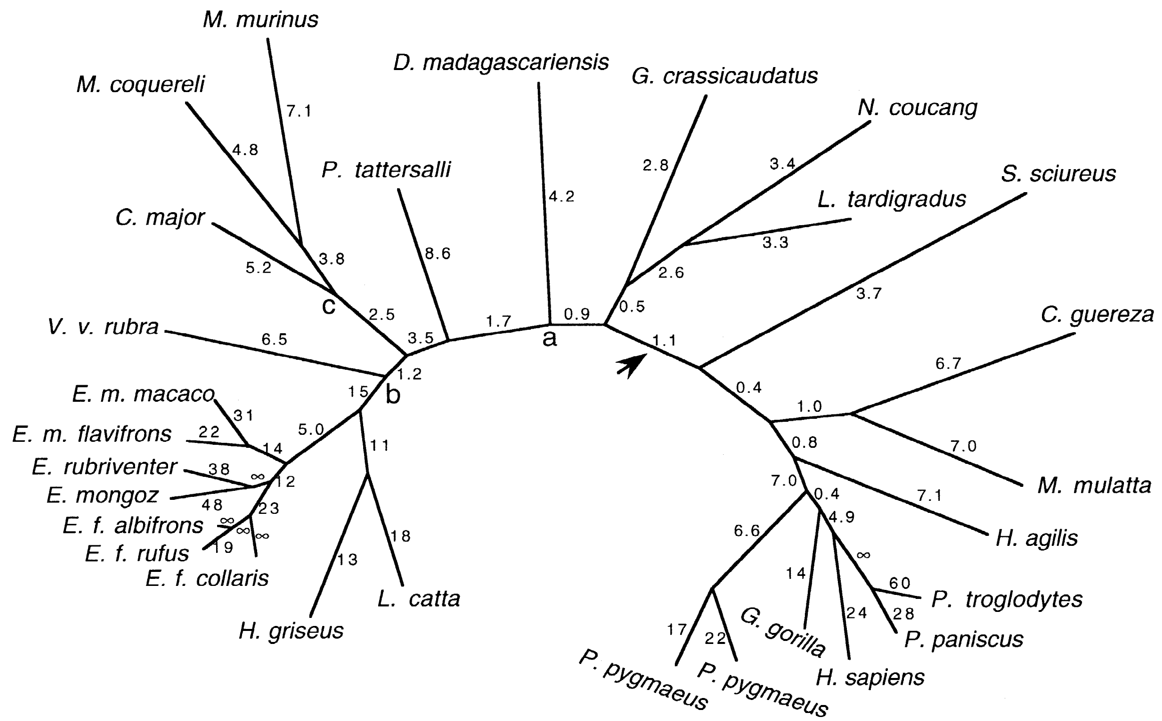

Fig. 2. The phylogenetic tree of the 28 primate species analyzed in this paper. The *arrow* indicates the root of the tree, although unrooted trees are used in the analyses in this paper. The position of *P. tattersalli* is uncertain, and it can be placed to cluster with the node *a* (as in the graph), *b,* or *c*. The mitochondrial cytochrome *b* genes are analyzed by maximum likelihood assuming the model of Kimura (1980) with a different ti/tv rate ratio ($\kappa = \alpha/\beta$ in the notation of Kimura) for each branch. Branches are drawn in proportion to their estimated lengths, measured as the number of nucleotide substitutions per site. Estimates of the ti/tv rate ratios (κ) under the model are shown along the branches.

Analytical Methods

Both pairwise and joint likelihood sequence comparisons are performed using two nucleotide substitution models. The first, referred to as "K80" (Kimura 1980), is a model of transition–transversion rate bias. The instantaneous substitution rate from nucleotide *i* to nucleotide *j* ($i \neq j$) under this model is

$$R_{ij} = \begin{cases} \alpha, & \text{for transition} \\ \beta, & \text{for transversion} \end{cases} \tag{1}$$

In a pairwise comparison, the sequence distance is defined as $d = (\alpha + 2\beta)t$, where *t* is the total time that separates the two sequences (twice the divergence time). The ti/tv rate ratio is defined as $\kappa = \alpha/\beta$. In a maximum-likelihood analysis of multiple sequences, the ti/tv rate ratio κ is often assumed to be constant among lineages with *d* used as the branch length. Both the sequence distance and branch length are measured by the expected number of nucleotide substitutions per site.

Let the proportions of sites in two sequences with transitional and transversional differences by *P* and *Q,* respectively. The sequence distance (*d*) and the transition/transversion rate ratio κ can be estimated as follows:

$$\hat{d} = -\frac{1}{2}\log\{1 - 2P - Q\} - \frac{1}{4}\log\{1 - 2Q\} \tag{2}$$
$$\hat{\kappa} = \frac{2\log\{1 - 2P - Q\}}{\log\{1 - 2Q\}} - 1$$

(Kimura 1980; Jukes 1987). These estimates, whenever obtainable, are also maximum-likelihood estimates.

The second model used in this paper, referred to as "HKY85" (Hasegawa et al. 1985), allows for both transition–transversion bias and unequal nucleotide frequencies. The substitution rate from nucleotide *i* to nucleotide *j* ($i \neq j$) is

$$R_{ij} = \begin{cases} \alpha\pi_j, & \text{for transition} \\ \beta\pi_j, & \text{for transversion} \end{cases} \tag{3}$$

where π_j is the frequency of nucleotide *j*. The expected number of nucleotide substitutions per site during time *t* (that is, the sequence distance or branch length) is

$$d = 2(\pi_T\pi_C + \pi_A\pi_G)\alpha t + 2\pi_Y\pi_R\beta t \tag{4}$$

where $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. The ti/tv rate ratio is again defined as $\kappa = \alpha/\beta$.

It should be noted that several definitions of the ti/tv rate ratio have been used in the literature. The definition we use in this paper (κ) is the ratio of instantaneous transition rate to instantaneous transversion rate ($\kappa = \alpha/\beta$) or the ratio of the number of transitions to the number of transversions after accounting for multiple substitutions at the same site ($\kappa = \alpha t/\beta t$); without any transition–transversion bias, $\kappa = 1$. It therefore differs from the ratio of the proportions of transitional to transversional differences observed between two sequences (*P*/*Q*) as used in many other studies (e.g., Wakeley 1994). The two measures are very different, as discussed later. Another definition of the ti/tv rate ratio is the ti/tv rate ratio averaged over the nucleotide frequencies (see, e.g., Wakeley 1996). This is equal to $\alpha/(2\beta)$ under the K80 model and $(\pi_T\pi_C + \pi_A\pi_G)\alpha/(\pi_Y\pi_R\beta)$ under HKY85.

In a maximum-likelihood analysis of multiple sequences, the ti/tv rate ratio κ under either the K80 or HKY85 model is estimated by numerical optimization, together with the branch lengths in the tree. A model that allows for different ti/tv rat ratios among lineages is also implemented. This model uses a separate κ parameter (and a branch length) for each branch in the tree and involves some modifications to

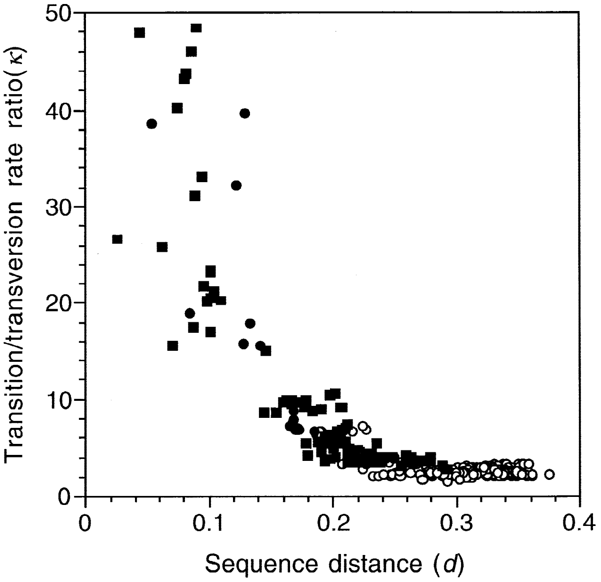

Fig. 3. Estimated transition/transversion rate ratio (κ) plotted as a function of the sequence distance from pairwise comparisons of mitochondrial cytochrome *b* genes. The 28 primate species in Fig. 2 are used. No transversional difference is observed ($Q = 0$) between the sequences for two eulemur species, *E.f. collaris* and *E.f. rufus,* yielding a κ of infinity, and thus this comparison is not shown in the graph. The model of Kimura (1980) is assumed, with a constant substitution rate for all sites. (■) Within the 15 lemuriform species; (●) within the seven hominoid species; (○) all other comparisons.

Table 2. Estimates of the transition/transversion rate ratio (κ) from cytochrome *b* genes during different stages of the stepwise addition algorithm[a]

Added species	K80	HKY85
A. For species in Fig. 4A		
1. Human (*H. sapiens*)		
2. Chimpanzee (*P. troglodytes*)	39.98	40.77
3. Gorilla (*G. gorilla*)	20.91	21.61
4. Orangutan (*P. pygmaeus*)	11.96	12.26
5. Gibbon (*H. agilis*)	9.57	9.84
6. Rhesus macaque (*M. mulatta*)	6.85	7.08
7. Squirrel monkey (*S. sciureus*)	5.14	5.29
8. Bushbaby (*G. crassicaudatus*)	4.20	4.31
9. Ring-tailed lemur (*L. catta*)	4.24	4.35
B. For species in Fig. 4B		
1. Eulemur (*E.f. collaris*)		
2. Eulemur (*E. mongoz*)	46.05	46.76
3. Ring-tailed lemur (*L. catta*)	12.43	12.71
4. Red ruffed lemur (*V. v. rubra*)	9.93	10.19
5. Dwarf lemur (*C. major*)	6.86	7.05
6. Aye aye (*D. madagascariensis*)	5.60	5.75
7. Bushbaby (*G. crassicaudatus*)	4.56	4.67
8. Squirrel monkey (*S. sciureus*)	4.04	4.14
9. Human (*H. sapiens*)	3.78	3.87

[a] Species in the trees in Figs. 4A (A) and 4B (B) are added into the tree in a stepwise manner, and a single κ parameter is assumed and estimated by maximum likelihood for all branches in the subtree at each stage of the algorithm.

the standard likelihood algorithm (Felsenstein 1981). The PAML program package (Yang 1997) is used for all analyses of this paper, and the DRAWTREE program in the PHYLIP package (Felsenstein 1997) is used to help draw the figures. Because the K80 and HKY85 models produced very similar estimates of the ti/tv rate ratios, results are often reported for the K80 model only. Patterns in estimates of the ti/tv rate ratios from three analyses are reported: (1) pairwise sequence comparison, (2) maximum likelihood in a stepwise addition algorithm for which more and more distant sequences are added in the tree, and (3) maximum likelihood under a model that allows for different ti/tv ratios among evolutionary lineages. We then explore possible causes for the observed patterns.

Results

Estimates of the Transition/Transversion Rate Ratio (κ) in Pairwise Sequence Comparisons

Figure 3 shows the estimated ti/tv rate ratios ($\kappa = \alpha/\beta$) plotted against the estimated sequence distances under the K80 model [Eq. (2)]. A total of 378 pairwise comparisons of the mitochondrial cytochrome *b* genes from the 28 species in Fig. 2 is illustrated. There is a strong negative correlation between estimates of *d* and κ. Because the variation at very low levels of sequence divergence may be due to sampling errors, there is an almost-perfect functional relationship between *d* and κ. Notably, the estimates of κ from comparisons within the hominoids are very similar to those from comparisons within the lemuriforms. This result is surprising with respect to the conclusions of Hasegawa et al. (1990) and suggests that species sampling may have an effect on the estimation of κ, as has been problematic for the estimation of κ using *P*/*Q*.

Maximum-Likelihood Estimates of the Transition/Transversion Rate Ratio (κ) in the Stepwise Addition Algorithm

We further examine how the maximum-likelihood estimate of the ti/tv rate ratio changes with the sampling of species, when the same ti/tv ratio is assumed for all lineages (branches) in the tree. An interesting approach for this purpose is through the stepwise addition algorithm. The algorithm starts from closely related species (such as different hominoids or different lemuriforms), with more and more divergent species added to the tree in a stepwise manner. Table 2A lists estimates of κ at each stage of the algorithm, which is used to construct the phylogeny in Fig. 4A, starting with the hominoids and ending with *Lemur catta*. The estimate of κ becomes smaller when more divergent sequences are added, and changes from about 40 for the human–chimpanzee comparison to 4 for the entire tree of nine species. The results therefore seem to suggest a low ti/tv ratio in the lemur and high ratios in the hominoids.

Table 2B lists estimates of κ when the stepwise addition algorithm is applied to construct the tree in Fig. 4B, in this case starting from the lemurid species, with the human sequence added last. Again, the estimate of κ

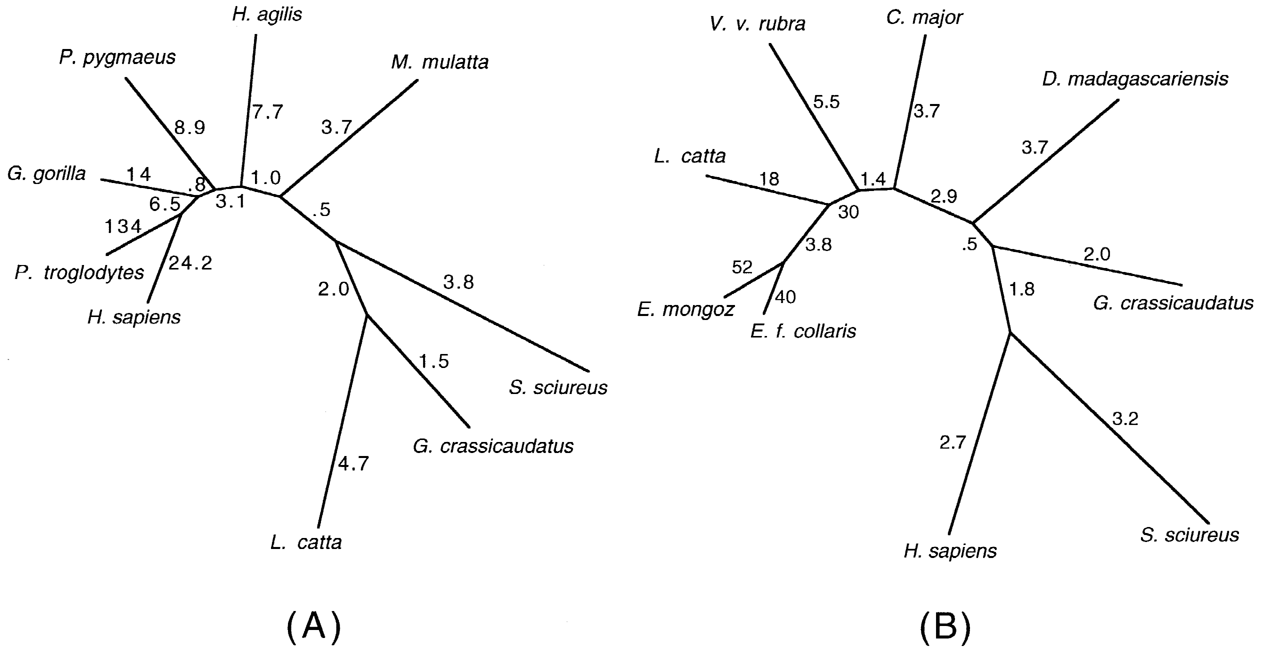

Fig. 4. Maximum-likelihood estimates of ti/tv (κ) for different branches in two phylogenetic trees of subsets of primate species in Fig. 2. The K80 model is assumed to analyze the mitochondrial cytochrome *b* genes, with a different ti/tv rate ratio assumed for each branch in the tree. **A** Many anthropoid species are used with two strepsirrhine species (*G. crassicaudatus* and *L. catta*) as outgroups. **B** Many strepsirrhine species are used with two anthropoid species (*S. sciureus* and *H. sapiens*) as outgroups.

becomes smaller with the addition of more divergent sequences, changing from 46 for the comparison of two *Eulemur* species to 3.8 for the entire tree. In contradiction to the results in Table 2A, the estimates seem to suggest that the ti/tv rate ratio may be high in the lemurs and low in the anthropoids.

Estimation of the Transition/Transversion Rate Ratios for Different Lineages

To test for the possibility that the ti/tv rate bias varies among the different primate lineages, we use a model that assumes different ti/tv rate ratios for different branches and apply it to the data in Figs. 4A and B. Estimates of the ti/tv rate ratios (κ) under this model are listed along branches of the trees. Since the model assuming a single ti/tv rate ratio is a special case of the model assuming variable ti/tv rate ratios, the likelihood-ratio test can be applied to compare the two models. The log-likelihood value under a model measures the model's fit to data. Because of the additional parameters, the model with variable ti/tv ratios among lineages is expected to fit the data better and have a higher log-likelihood value than the model assuming a single ti/tv ratio. However, the improvement in the log-likelihood value by relaxing the assumption of constancy of the ti/tv ratio will not be statistically significant if the single-ratio model provides an adequate fit to the data. For the tree in Fig. 4A, the log-likelihood value under the model with variable ti/tv ratios among branches is $l_1 = -6548.77$, while the model assuming a single ratio for the entire tree has log-likelihood value $l_0 = -6464.75$. The variable-ratio model involves 30 parameters, with 2 parameters (length d and ti/tv ratio κ) for each of the 15 branches in the tree, while the constant-ratio model has 16 parameters (15 branch lengths and 1 κ). Twice the log-likelihood difference, $2\Delta l = 2(l_1 - l_0) = 168.04$, should be compared with $\chi^2 = 29.1$ with df = 14 (30 − 16) at the 1% significance level. The two models are significantly different, with different ti/tv ratios among branches being preferred.

The species included in the tree in Fig. 4A are similar to those analyzed by Hasegawa et al. (1990), although those authors used a different segment of the mitochondrial genome. Even so, the estimates of ti/tv rate ratios in Fig. 4A agree with Hasegawa et al.'s (1990) results. The ti/tv ratios in the hominoids, especially in humans and chimpanzees, appear to be much higher than those in the more-divergent outgroup species such as the lemur and the Old World and New World monkeys.

For the tree in Fig. 4B, the log-likelihood values under the variable-ratio and constant-ratio models are $l_1 = -6605.68$ and $l_0 = -6667.57$, respectively, with $2\Delta l = 2(l_1 - l_0) = 123.78$. The results again indicate that ti/tv rate ratios are significantly different among branches. However, estimates of the ti/tv ratios for branches of this tree under the variable-ratio model show a reversed pattern to that in Fig. 4A, with ti/tv rate ratios appearing to be much higher in the lemurs than in the hominoids (human).

The model of variable ti/tv rate ratios among lineages is also applied to the tree of all 28 primate species in Fig. 2, and the estimates of the ti/tv ratios are shown along the branches of the tree. The log-likelihood value under this model is $l_1 = -14361.36$. The model assuming a constant ti/tv rate ratio among branches has log-likelihood value $l_0 = -14584.19$, with $\hat{\kappa} = 5.26$. The phylogenetic tree has 53 branches, and so the numbers of parameters in the two models differ by 52. Comparison of $2\Delta l = 2(l_1 - l_0) = 445.66$, with $\chi^2 = 78.6$ with df $= 52$ at the 1% significance level suggests that the two models are significantly different, again with a strong preference for variable ti/tv rate ratios among lineages.

Exploring Possible Causes for the Observed Patterns

If the ti/tv rate ratio in the cytochrome *b* gene has been constant during the primate evolution, and if the estimation procedures adequately corrects for multiple substitutions, estimates of κ should be independent of the sequence divergence. The strong negative correlation between estimates of *d* and κ (Fig. 3), and the dependence of the maximum likelihood estimate of κ on the species included in the sample (Table 2) therefore suggest that either the ti/tv rate ratios are not constant over time, or some assumptions involved in the model are unrealistic, or the estimation procedures involve substantial biases. Several of the model's assumptions are examined below. We note that the relative nucleotide frequencies are quite homogeneous among the taxa examined and, thus, do not appear to have varied much during the primate evolution. Although it is true that the four nucleotides have unequal frequencies (26, 33, 29, and 12% for T, C, A, and G, respectively), use of the HKY85 model to account for the nucleotide frequency bias produced very similar results to K80 (results not shown), suggesting that unequal nucleotide frequencies do not account for the observed pattern. A similar argument suggests that codon usage bias is unlikely to explain the patterns in the estimated ti/tv rate ratios.

Variation of Evolutionary Rates Among Codon Positions or Nucleotide Sites. The nucleotide substitution rates at the three codon positions in the cytochrome *b* genes are in the proportion 1:0.26:10.23, estimated by a likelihood model that accounts for different rates at the three positions (Yang 1996). The base frequencies and apparent ti/tv rate biases are also quite different at the three codon positions (Yoder et al. 1996b). To see whether the observed pattern is due to different evolutionary dynamics at the codon positions, pairwise comparisons are also performed for the three positions separately. The same pattern of negative correlation between *d* and κ is seen at all codon positions, although the pattern is clearest at the third position, followed by the first and second. The second positions are not very variable, and estimates obtained from them therefore involve large sampling errors.

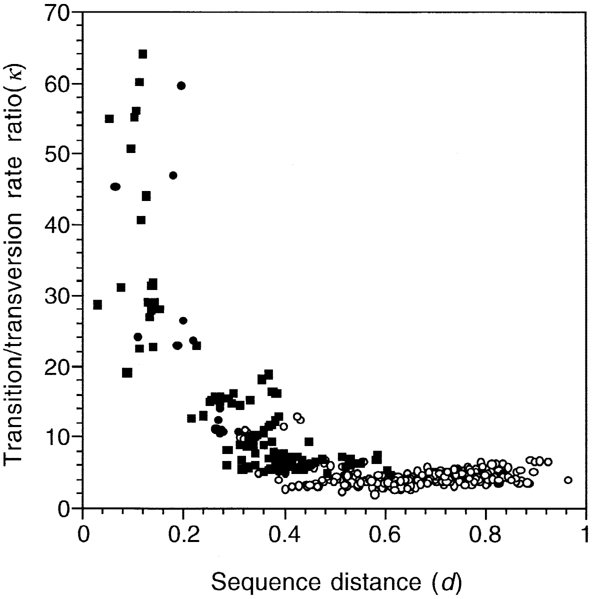

Fig. 5. Estimated ti/tv rate ratio (κ) plotted as a function of the sequence distance (*d*) from pairwise comparisons of mitochondrial cytochrome *b* genes. The model of Kimura (1980) is used, with substitution rates among sites assumed to be gamma distributed, and the gamma shape parameter set at 0.32, the maximum-likelihood estimate for all the 28 species in Fig. 2. (■) Within the 15 lemuriform species; (●) within the seven hominoid species; (○) all other comparisons.

The substitution rates in the cytochrome *b* genes, especially at the second and first codon positions, are highly variable among sites. When the HKY85 model is applied to the entire data set containing all species in Fig. 2, with substitution rates approximated by a discrete-gamma model with five rate categories (Yang 1994), the shape parameter of the distribution is estimated to be 0.32, indicating severe rate variation among sites. This estimate is then used to estimate the sequence distance (*d*) and the ti/tv rate ratio in pairwise sequence comparisons under the K80 model (Jin and Nei 1990). The results are shown in Fig. 5. Estimates of both the sequence distance and the ti/tv rate ratio under this model are considerably larger than the corresponding estimates assuming a constant rate for all sites (Fig. 3). This result is compatible with the previous observation that ignoring rate variation among sites causes underestimation of the sequence distance and the ti/tv rate ratio (Wakeley 1994; Yang et al. 1994). Nonetheless, estimates of *d* and κ again show the same negative correlation, regardless of the model employed (Figs. 3 and 5).

Bias in the Estimation Procedures. One may suspect that the pattern in Fig. 3 may be due to biases in the correction formula [Eq. (2)]. If the ti/tv ratio is overestimated at a low sequence divergence and/or underestimated at high sequence divergence, a negative correlation between estimates of *d* and κ will be generated. The pattern in Fig. 3 would not have been surprising if we had used the ratio *P*/*Q* in our analysis, that is, the ratio of

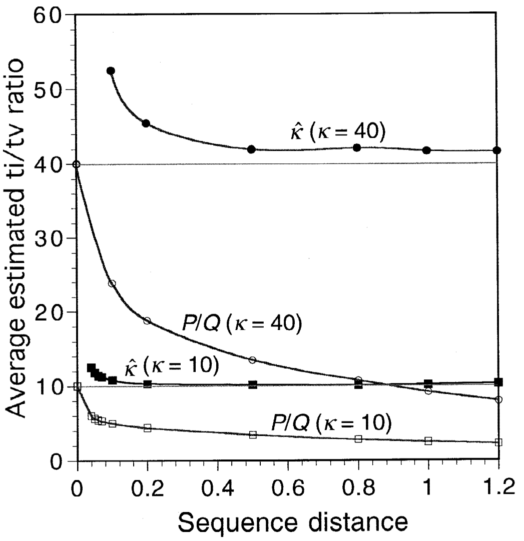

Fig. 6. Average, among 1000 simulated replicates, of pairwise estimates of the ti/tv rate ratio by Eq. 2 ($\hat{\kappa}$ for *filled symbols*) and by P/Q (*open symbols*) as functions of the true sequence distance. The K80 model is used to simulate data sets of pairs of sequences. The true ti/tv rate ratio is $\kappa = 10$ (squares) and 40 (circles). The sequence length is 1000. Simulation results at very low sequence divergence ($d < 0.04$ for $\kappa = 10$ and $d < 0.1$ for $\kappa = 40$) are not presented because Eq. (2) is sometimes inapplicable and excluding the inapplicable cases would bias results. Inapplicable cases also occur at a very high sequence divergence.

the proportions of transitional and transversional differences between two sequences. A strong negative correlation between P/Q and d indeed exists for these data (Fig. 1D), and saturation of transitions at a high sequence divergence is clearly a very important cause. However, the bias in the estimate of κ from the method of Kimura (1980) and Jukes (1987), which corrects for multiple hits, is less clear.

Computer simulations are thus performed to examine the bias in the estimate of the ti/tv ratio by both Kimura's correction [Eq. (2)] and P/Q. Figure 6 shows estimates of κ in pairwise sequence comparisons when the true κ is fixed at either 10 or 40. The performance of the two measures are very different. The ratio P/Q always underestimates the ti/tv rate ratio and decreases very quickly with the increase in the sequence divergence, approaching 1/2 irrespective of the true ti/tv rate ratio. Saturation of transitions is clearly the major reason for the negative correlation between P/Q and d. However, the bias in the estimate of κ by Eq. (2) appears always to be positive; that is, Kimura's formula always overcorrects for multiple transitions at the same site and thus overestimates the ti/tv rate ratio. The overestimation is negligible at a high sequence divergence but is substantial at a low divergence. Note that simulation results are not obtained for very low sequence divergences in Fig. 6, although they are commonplace in our analysis of the real data (Fig. 3), indicating that the overestimation may be more serious than the results in Fig. 6 suggest.

Simulations were also performed to examine the bias in the likelihood method applied to multiple sequences, which should be even more tolerant of multiple substitutions than pairwise comparison. In one simulation, sequences of 1000 sites are generated from an unrooted model tree of five species. All seven branch lengths are set at 0.5 substitution per site, so that the tree length (the total number of substitutions per site on the tree) is 3.5. The K80 model is assumed, with κ fixed at 50. At this level of sequence divergence, the pairwise distance formula [Eq. (2)] is inapplicable in almost every simulated data set. The average of the maximum-likelihood estimate of κ over 500 replicates is 50.8, with a very slight positive bias. Even at such a high sequence divergence, saturation of transitions is not a problem. When the seven branches of the tree are all shortened to 0.02, so that the tree length is 0.14, the average of the estimate of κ is 66.7, with a much larger bias. Thus, both pairwise comparison and likelihood joint analysis tend to overcorrect for multiple transitions at the same site and overestimate κ, particularly at a low sequence divergence. The bias is also greater when the true ti/tv ratio (κ) is higher and when the sequences are shorter (results not shown).

Overestimation of κ at a low sequence divergence has clearly contributed to the negative correlation between estimates of d and κ seen in Fig. 3. However, the bias of the estimation method cannot fully account for the pattern. If the observed pattern were due to the bias in the method, reliable estimates of κ would be those at high levels of sequence divergence, according to the simulation results discussed above, that is, a κ of about 2–3 (Fig. 3). With such a low ti/tv rate ratio, however, the bias in Kimura's correction would be very small (Fig. 6) and could not possibly create such large estimates of κ as 40 (Fig. 3).

Variable ti/tv Rate Ratios Among Evolutionary Lineages. Results in Table 2 and, in particular, the likelihood ratio tests of Figs. 2 and 4A and B suggest that the ti/tv rate ratios may be variable among evolutionary lineages. If the transition and transversion rates are indeed variable among lineages, the pattern in Fig. 3 may be explained by estimates of the branch lengths and ti/tv rate ratios for branches shown in Fig. 2.

Suppose that the transition and transversion rates during time t_1 are α_1 and β_1, respectively, and the transition and transversion rates during time t_2 are α_2 and β_2. It is easy to show (see Appendix) that the ti/tv rate ratio during the entire time period $(t_1 + t_2)$ is $(\alpha_1 t_1 + \alpha_2 t_2)/(\beta_1 t_1 + \beta_2 t_2)$. This result applies to both pairwise comparison or a branch in the phylogenetic tree in a likelihood analysis. If the real ti/tv ratio varies over time but a constant ratio is assumed for estimation, the resulting estimate will be an ''average'' of the ratios over the entire time period. This result can be used to calculate the expected ti/tv rate ratios for pairwise comparisons using the estimates in Fig. 2. For example, estimates of αt along the 14 branches in the tree in Fig. 2 linking the ring-tailed lemur to human are 0.0573, 0.0292, 0.0208, 0.0057, 0.0150, 0.0252, 0.0091, 0.0203, 0.0074, 0.0064, 0.0148, 0.0022, 0.0107, and 0.0638, with a sum of 0.2878 transitions per site. The corresponding estimates of βt are 0.0031, 0.0027, 0.0014, 0.0048, 0.0042, 0.0151, 0.0103, 0.0180, 0.0200, 0.0084, 0.0021, 0.0050, 0.0022, and 0.0027, with the sum to be 0.1000 (which means 0.2 transversions per site between the two sequences). The estimate of $\kappa = \alpha/\beta$ for the pairwise comparison (assuming a constant ti/tv rate ratio) is expected to be $0.2878/0.1000 = 2.88$. This value is close to 2.56, the estimate from Eq. (2) for the pairwise comparison of the two sequences. The average of the ti/tv ratios (κ) across the 14 branches (Fig. 2) is 6.41.

Furthermore, estimates of the ti/tv ratios among branches shown in Fig. 2 suggest the unexpected conclusion that the ti/tv ratios are small near the root of the primate phylogeny. In particular, several long interior branches near the root are all characterized by low ti/tv rate ratios. This pattern, together with the overestimation of the ti/tv rate ratio at low sequence divergence, will create a strong negative correlation between estimates of the ti/tv rate ratio and sequence distance.

Discussion

The dependence of the estimate of the ti/tv rate ratio on the sampled species in either pairwise comparison or joint likelihood analysis is more complex than has been previously considered. Likelihood-ratio tests suggest that the ti/tv rate ratios may be variable among evolutionary lineages. This variation, together with the overestimation of the ti/tv rate ratio at low sequence divergence, appears sufficient to account for the observed patterns. Even so, the hypothesis of changing ti/tv rate ratios, as either a substitutional or mutational dynamic, is difficult to substantiate. Maximum-likelihood estimates of synonymous/nonsynonymous rate ratios for branches in the tree in Fig. 2 (obtained from an extension of the model of Goldman and Yang 1994) are not correlated with estimates of the ti/tv rate ratios, suggesting that variation of selectional constraints is probably unlikely to explain the variable ti/tv ratios among lineages. It is also unlikely that mutation rates could have varied so drastically on the phylogenetic scale represented by the primates. Although the patterns in Figs. 3 and 5 appear compatible with the expectation of saturation at high sequence divergence and/or even more substantial overcorrection at a low sequence divergence, such biases in the estimation procedure are unlikely to explain the observed pattern which is at a much greater scale. Nucleotide frequency bias and variable evolutionary rates among sites and at the three codon positions can be ruled out as possible explanations. It seems, therefore, that the pattern is due to some unknown factor that has not been considered in our models. The models (K80 or HKY85) used in this paper are simple and may not capture the complexities of cytochrome *b* evolution. Moreover, it is not clear which aspects of the evolutionary process that have not been accounted for in the models might have caused the observed patterns. This emphasizes the need for continued exploration of the multiple factors that affect sequence evolution.

The overcorrection of multiple transitions at the same site by Kimura's formula (or the maximum-likelihood method in general) at a low sequence divergence is an unexpected result and may have important implications. Most significantly, it suggests that assessing rate parameters from recently diverged taxa may bias conclusions in favor of overestimation of transition rates.

Acknowledgments. We Thank Joe Felsenstein, Paul Lewis, Magnus Nordborg, Bruce Rannala, Jeff Thorne, and John Wakeley for helpful discussions and comments. A.D.Y. was supported by NSF Grant DEB-9303313.

References

Anderson SA, Bankier T, Barrell BG, et al. (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates, tempo and mode of evolution. J Mol Evol 18:225–239

Collura RV, Stewart C-B (1995) Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature 378:485–489

Curtis SE, Clegg MT (1984) Molecular evolution of chloroplast DNA sequences. Mol Biol Evol 1:291–301

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

Felsenstein J (1997) PHYLIP (phylogeny inference package), Version 4.0. University of Washington, Seattle

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

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

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

Hasegawa M, Kishino H, Hayasaga K, Horai S (1990) Mitochondrial DNA evolution in primates: transition rate has been extremely low in the lemur. J Mol Evol 31:113–121

Hayasaga K, Gojobori T, Horai S (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol 5:626–644

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92: 532–536

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

Jukes TH (1987) Transitions, transversions, and the molecular clock. J Mol Evol 26:87–98

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

Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. Annu Rev Ecol Syst 18:269–292

Purvis A, Bromham L (1997) Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. J Mol Evol 44:112–119

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

Wakeley J (1996) The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. TREE 11:158–163

Wills C (1995) When did Eve live? An evolutionary detective story. Evolution 49:593–607

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

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

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

Yoder AD (1997) Back to the future: a synthesis of strepsirrhine systematics. Evol Anthropol 6:11–22

Yoder AD, Irwin JA (1999) Phylogeny of the Lemuridae: Effects of taxon and character sampling on resolution of species relationships with Eulemur. Cladistics (in press)

Yoder AD, Cartmill M, Ruvolo M, Smith K, Vilgalys R (1996a) Ancient single origin for malagasy primates. Proc Natl Acad Sci USA 93:5122–5126

Yoder AD, Vilgalys R, Ruvolo M (1996b) Molecular evolutionary dynamics of cytochrome b in strepsirrhine primates: The phylogenetic significance of third position transversions. Mol Biol Evol 13:1339–1350

Appendix. Estimation of the Transition/Transversion Rate Ratio When the Ratio Varies Over Time

The substitution rate matrix $R = \{R_{ij}\}$ [Eq. (3)] under the HKY85 model (Hasegawa et al. 1985) can be written as

$$R = \begin{bmatrix} -(\alpha\pi_C + \beta\pi_R) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & -(\alpha\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha\pi_G + \beta\pi_Y) & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & -(\alpha\pi_A + \beta\pi_Y) \end{bmatrix} \tag{A1}$$

where $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$, and the nucleotides are ordered T, C, A, and G. The diagonals are determined by the mathematical requirement that sums of rows of the matrix are all zero. The eigenvalues and eigenvectors of R are obtained by Hasegawa et al. (1985), so that the matrix can be diagonalized as follows:

$$R = U\,\Lambda\,U^{-1} \tag{A2}$$

$\Lambda = \mathrm{diag}\{0, -\beta, -(\alpha\pi_R + \beta\pi_Y), -(\alpha\pi_Y + \beta\pi_R)\}$ is a diagonal matrix with the diagonals to be the eigenvalues of R and the off-diagonals to be zero.

$$U = \begin{bmatrix} 1 & 1/\pi_Y & 0 & \pi_C/\pi_Y \\ 1 & 1/\pi_Y & 0 & -\pi_T/\pi_Y \\ 1 & -1/\pi_R & \pi_G/\pi_R & 0 \\ 1 & -1/\pi_R & -\pi_A/\pi_R & 0 \end{bmatrix} \tag{A3}$$

is a matrix of the corresponding right eigenvectors, and

$$U^{-1} = \begin{bmatrix} \pi_T & \pi_C & \pi_A & \pi_G \\ \pi_T\pi_R & \pi_C\pi_R & -\pi_A\pi_Y & -\pi_G\pi_Y \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \end{bmatrix} \tag{A4}$$

is the inverse of U; that is, $U\,U^{-1} = U^{-1}\,U = \mathrm{I}$, the identity matrix. Note that matrices U and U^{-1} are functions of the nucleotide frequencies only, independent of the transition and transversion rates, α and β. We assume that nucleotide frequencies do not change over time, and only the transition and transversion rates are variable.

Since time and rate are confounded, the transition probability matrix over time t is a function of αt and βt only. Let this be $M(\alpha t, \beta t)$. Thus

$$\begin{aligned} M(\alpha t, \beta t) &= \exp(Rt) = U\exp(\Lambda t)\,U^{-1} \\ &= U\,\mathrm{diag}\{1, \exp\{-\beta t\}, \exp\{-(\alpha\pi_R + \beta\pi_Y)t\}, \\ &\qquad \exp\{-(\alpha\pi_Y + \beta\pi_R)t\}\}\,U^{-1} \end{aligned} \tag{A5}$$

Suppose that the transition and transversion rates are α_1 and β_1 during time t_1, and α_2 and β_2 during time t_2. The transition probability matrix over the entire time period $(t_1 + t_2)$ is easily shown to be

$$M(\alpha_1 t_1, \beta_1 t_1) \times M(\alpha_2 t_2, \beta_2 t_2) = M(\alpha_1 t_1 + \alpha_2 t_2, \beta_1 t_1 + \beta_2 t_2) \tag{A6}$$

The statistical behavior of the sequence data when the transition and transversion rates are α_1 and β_1 during time t_1 and α_2 and β_2 during time t_2 is determined by two parameters only: $(\alpha_1 t_1 + \alpha_2 t_2)$ and $(\beta_1 t_1 + \beta_2 t_2)$. If a constant ti/tv rate bias is assumed while in fact the ratio is variable, we will obtain the correct estimate of the sequence distance for the entire period, but the ti/tv rate ratio will be estimated as $(\alpha_1 t_1 + \alpha_2 t_2)/(\beta_1 t_1 + \beta_2 t_2)$. The result applies to the K80 model too since K80 is a special case of HKY85.