40 Pesole, G. *et al.* (1991) **The branching order of mammals: Phylogenetic trees inferred from nuclear and mitochondrial molecular data,** *J. Mol. Evol.* 33, 537–542

41 Doyle, J.J. (1992) **Gene trees and species trees: Molecular systematics as one-character taxonomy,** *Syst. Bot.* 17, 144–163

42 Hillis, D.M. *et al.* (1991) **Evidence for biased gene conversion in concerted evolution of ribosomal DNA,** *Science* 251, 308–310

43 Lockhart, P.J. *et al.* (1994) **Recovering evolutionary trees under a more realistic model of sequence evolution,** *Mol. Biol. Evol.* 11, 605–612

44 Huelsenbeck, J.P. and Hillis, D.M. (1993) **Success of phylogenetic methods in the four-taxon case,** *Syst. Biol.* 42, 247–264

45 Hendy, M. and Penny, D. (1989) **A framework for the quantitative study of evolutionary trees,** *Syst. Zool.* 38, 297–309

The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance

John Wakeley

Estimates of transition bias provide insight into the process of nucleotide substitution, and are required in some commonly used phylogenetic methods. Transitions are favored over transversions among spontaneous mutations, and the direction and strength of selection on proteins and RNA appears to depend on mutation type. As the complexity of the nucleotide-substitution process has become apparent, problems with classical methods of estimating transition bias have been recognized. These problems arise because there is a fundamental difference between ratios of numbers of differences among sequences and ratios of rates, and because classical methods are not easily generalized. Several new methods are now available.

John Wakeley is at the Dept of Biological Sciences, Nelson Biological Labs, PO Box 1059, Busch Campus, Rutgers University, Piscataway, NJ 08855-1059, USA.

The four bases of DNA are classified by their structures into purines and pyrimidines (see Fig. 1). Nucleotide substitutions within each structural class (transitions) occur with greater frequency than changes between structural classes (transversions). This phenomenon is called transition bias and was discovered when the first comparisons of molecular sequences were made[1,2]. It is now recognized as a general property of DNA-sequence evolution, having been observed in nuclear, mitochondrial and chloroplast DNA, and in prokaryotes, eukaryotes and viruses. Transition bias has been found in the DNA sequences of pseudogenes and functional genes[3,4], in transfer RNA (tRNA) (Ref. 5), in ribosomal RNA (rRNA) (Ref. 6) and in the functional but non-coding mitochondrial control region[7]. While always present, transition bias appears more pronounced in animal mitochondrial[8] than in nuclear[3,4] or chloroplast[9] DNA.

Studies of transition bias are valuable for two reasons. First, estimates of the pattern of nucleotide substitution are important to our understanding of DNA-sequence evolution. Knowledge of transition bias facilitates inferences about mutational patterns and about the type and strength of natural selection. Second, reliable estimates of transition bias are important to evolutionary-distance correction methods (see Goldstein and Pollock[10] for a recent example). Changes in estimates of transition bias can substantially alter distance corrections[11]. Corrected distances are used as input into phylogenetic-tree-building algorithms and to estimate divergence times.

Models of nucleotide substitution

One cannot rightly discuss transition bias without invoking models of nucleotide substitution, and many of these have been proposed. The definition of transition bias adopted here is the one used by most workers and is easily applied to all substitution models. Transition bias is measured by the ratio of the overall rate of transitions (*TI*) to the overall rate of transversions (*TV*). This quantity is referred to here as the *TI*:*TV* rate ratio.

Jukes and Cantor[12] introduced the first and simplest substitution model. Since all changes are considered equally likely, this model contains no transition bias. However, the *TI*:*TV* rate ratio is 1:2, simply because there are twice as many transversions as transitions (see Fig. 1). Transition bias is indicated when the *TI*:*TV* rate ratio is greater than 1:2, and many substitution models incorporate this possibility explicitly. Kimura's[13] two-parameter model is a model of only transition bias. Transitions happen at rate α and transversions at rate β, so the *TI*:*TV* rate ratio is α:(2β). When $\alpha = \beta$, this model reduces to that of Jukes and Cantor, and when $\alpha > \beta$, there is transition bias.

As the complexities of DNA-sequence change were elucidated, new models were introduced in an attempt to capture something more of the reality of nucleotide substitution. Almost all of these models include transition bias. Box 1 shows three currently used models and gives expressions for the *TI*:*TV* rate ratio for each.

The causes of transition bias

In the same year as their proposal of the structure of DNA, Watson and Crick[14] suggested a mechanism of point mutation that immediately favors transitions. Their idea was that while the maintenance of the double helix demanded that purines always paired with pyrimidines and vice versa, mutations could occur if disfavored tautomeric forms of the

PII: S0169-5347(96)10009-4

four bases were misincorporated during DNA replication to form A–C and G–T transition mispairs. In 1976, Topal and Fresco[15] elaborated a scheme whereby transversions could also occur via purine–purine mispairs by the misincorporation of a disfavored tautomer together with a rotation of the template base relative to its sugar. Transition mispairs were thought to be more likely than transversion mispairs because they did not require a change in conformation.

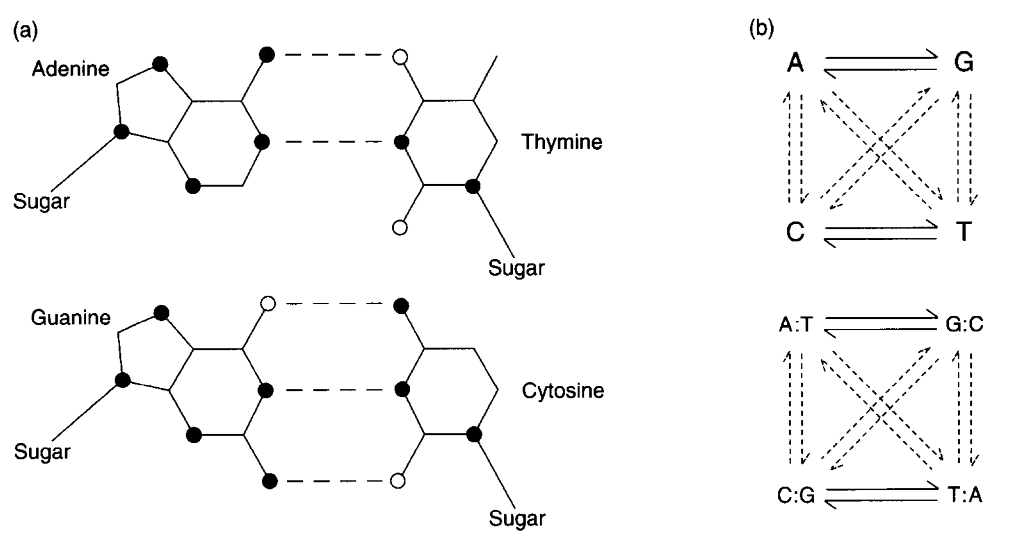


Fig. 1. The four nucleotides fall into two structural classes: purines [adenine (A) and guanine (G)] and pyrimidines [cytosine (C) and thymine (T)]. As shown in (a), A pairs with T and G pairs with C in the DNA double helix. The open circles in (a) represent oxygen atoms, the shaded circles are nitrogen atoms, and all other vertices are carbons. Hydrogen atoms are not shown, but the hydrogen bonds that form between paired bases are indicated by dashed lines. Because there are four nucleotides, the 12 changes shown in (b) are possible. The phenomenon that substitutions of one purine for another purine or one pyrimidine for another pyrimidine (transitions) occur with greater frequency than substitutions between purines and pyrimidines (transversions) is known as transition bias. This is emphasized in (b) by the use of solid arrows for transitions and dashed arrows for transversions. The lower part of (b) reminds us that substitutions actually occur by the replacement of base pairs in the DNA helix.

The incorporation of disfavored tautomers of bases into a growing DNA strand has not subsequently been observed to play a significant role in mutation[16]. Instead, it appears that the standard forms of the bases are misincorporated directly, and in many different pairings. These include purine–purine and pyrimidine–pyrimidine pairs, and the orientation of mispairs can differ greatly from the standard helical geometry[17]. However, the free energies of pairings between the template base and incoming deoxynucleoside triphosphates as well as the pressure to maintain the Watson–Crick structure do influence the rates of misincorporation; consequently, G–T, G–A, and A–C are generally the most frequent mispairs[16]. Thus, the bias towards transitions starts at the level of mutation.

Population-level processes can also influence transition bias. For neutral mutations (i.e. those with negligible selective advantage or disadvantage), fixation occurs by random genetic drift at a rate equal to the rate of mutation. On the other hand, the rate of substitution of selected mutations depends on the type and strength of selection. For example, in protein-coding sequences, replacement changes (ones which alter the amino acid sequence of the peptide) are generally fixed at a different (usually lower) rate than synonymous changes. If there is some correlation between the type of mutation (either transition or transversion) and the type and/or strength of selection, then selective forces can alter the transition bias.

There appear to be such correlations in both protein-coding and RNA-coding sequences, but the situation is far from simple. Among single-step substitutions in the universal genetic code at third codon positions, only about 3% of transitions cause amino acid replacements, compared with 41% for transversions. For the mammalian mitochondrial genetic code, these figures are 0% and 43%. Thus, Crozier and Crozier[18] recently cited conservation of amino acid sequence as a probable cause of an excess of transitions at third positions among some highly diverged sequences.

At first and second positions, transitions and transversions appear equally favored or disfavored since nearly all changes at these positions cause amino acid replacements. However, Vogel and Kopun[19] studied only replacement changes and showed that transition mutations tend to cause changes that conserve the chemical properties of amino acids. If selection acts to conserve the chemical nature of proteins, transitions will be favored over transversions. This was recently illustrated by Naylor *et al.*[20], who showed that the maintenance of membrane-spanning hydrophobic domains in mitochondrial proteins can constrain second codon positions to be either C or T; transitions may occur freely at these sites but transversions are deleterious.

In rRNA-coding and tRNA-coding regions, selection probably acts to maintain secondary structure in the product. In RNA stems, which are by definition paired regions, G–U is the most frequently observed non-canonical pair and is believed to maintain stem structures[21]. Thus, as in the DNA itself, selection for base pairing in RNA stems may favor transitions. Vawter and Brown[22] cite this as the reason why they observe consistent transition bias in stem regions of small subunit rRNA but not elsewhere in the molecule. Crozier and Crozier[18] put forth a similar idea: they propose that an excess of transitions observed in small subunit rRNA results from selection acting to conserve nucleotide size.

Classical methods of estimating transition bias

The first introduced and still most commonly used method of estimating transition bias is to count the differences among some sequences, then simply divide the number of transitions by the number of transversions. Given that there are some underlying rates for transitions and transversions, it is hoped that the numbers of each will reflect these rates and that their ratio will estimate the *TI*:*TV* rate ratio. Two ways of counting changes were adopted early and persist today. One is to take pairs of sequences in the sample and compare them (for example, see Ref. 8). The other is to construct a tree relating all the sequences in a sample, then use parsimony to infer the minimum number of changes that must have occurred at each site[23]. Figure 2 shows worked examples of these two methods.

Regardless of how changes are counted, the ratio of the numbers of transition and transversion differences is an accurate estimator of the *TI*:*TV* rate ratio only under certain limited conditions. This is illustrated below for the pairwise method of counting changes and Kimura's substitution model, but the points discussed hold also for the parsimony method and for other models of nucleotide substitution.

Problems in estimating transition bias

The probabilities of observing each possible pair of nucleotides at a site in two sequences that descend from a common ancestral sequence have been derived for many substitution models. For Kimura's two-parameter model, it is simplest to combine these into just two categories: *P* for transitions and *Q* for transversions. Kimura[13] found that:

$$P(t)=\frac{1}{4}-\frac{1}{2}e^{-4(\alpha+\beta)t}+\frac{1}{4}e^{-8\beta t} \qquad (1)$$

and

$$Q(t)=\frac{1}{2}-\frac{1}{2}e^{-8\beta t} \qquad (2)$$

are the probabilities of observing a transition or a transversion at a site in two sequences separated from their common ancestor by a length of time, *t*.

Any particular site will either show a transition or a transversion, or it will be identical in both sequences. Between two sequences *n* sites long, we expect to observe $n\times P(t)$ transition changes and $n\times Q(t)$ transversion changes. If *n* is large, the observed numbers of transitions and transversions should be close to these expectations, and their ratio should be approximately $P(t)$:$Q(t)$.

Unfortunately, $P(t)$:$Q(t)$ will not always be close to the true *TI*:*TV* rate ratio, α:(2β). When *t* is near zero, $P(t)$ is approximately $2\alpha t$ and $Q(t)$ is approximately $4\beta t$. In this case, if *n* is large enough, $P(t)$:$Q(t)$ should be nearly α:(2β). However, when *t* is very great, $P(t)$ approaches 1/4 and $Q(t)$ approaches 1/2, independent of α and β (see Fig. 3). At this extreme, the *TI*:*TV* rate ratio is estimated to be 1:2, indicating no transition bias. The ratio of the probability of transition to the probability of transversion decreases with time for all substitution models that include transition bias. Thus, the pairwise method and the tree-based parsimony method underestimate the *TI*:*TV* rate ratio when sequences are substantially diverged.

The obvious strategy of using only very recently diverged sequences may not be ideal either, because of the discrete nature of sequence data. For example, with strong transition bias, the first few changes between two sequences are likely to be transitions and the resulting *TI*:*TV* rate ratio estimate will be infinity (see Fig. 2). The ratio of the observed numbers of transitions and transversions is a biased estimate of the *TI*:*TV* rate ratio when a finite number of sites is examined, even if $P(t)$:$Q(t)$ is close to α:(2β). The effect of this is strongest when few sites are examined in sequences with high transition bias, and the variances of such estimates can be huge[24].

A third problem in quantifying transition bias is the existence of substitution-rate variation among nucleotide sites, a phenomenon that appears to be as fundamental as transition bias in molecular evolution. If some sites in a sequence evolve rapidly and others evolve slowly, the faster sites will experience multiple substitutions soon after divergence. This biases estimates of the *TI*:*TV* rate ratio towards the equilibrium value, which, as discussed above, is likely to be much smaller than the actual bias. When substitution-rate variation exists among sites but is ignored, transition bias is underestimated[24]. This is true not only of the pairwise and tree-based parsimony methods, but of maximum likelihood as well[25].

Box 1. Models of nucleotide substitution

Kimura's model[13]

i\j	A	G	C	T
A	*	α	β	β
G	α	*	β	β
C	β	β	*	α
T	β	β	α	*

Hasegawa *et al.*'s model[34]

i\j	A	G	C	T
A	*	$\alpha\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$\alpha\pi_A$	*	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	*	$\alpha\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$\alpha\pi_C$	*

Felsenstein's model[31]

i\j	A	G	C	T
A	*	$(1+\frac{\kappa}{\pi_R})\pi_G$	π_C	π_T
G	$(1+\frac{\kappa}{\pi_R})\pi_A$	*	π_C	π_T
C	π_A	π_G	*	$(1+\frac{\kappa}{\pi_Y})\pi_T$
T	π_A	π_G	$(1+\frac{\kappa}{\pi_Y})\pi_C$	*

The entries, λ_{ij}, of each matrix above are the probabilities of nucleotide *i* being replaced by nucleotide *j* in one time unit (for example, one generation or one year). Many authors use a continuous-time approximation in which the λ_{ij} are assumed to be instantaneous rates of nucleotide substitution from *i* to *j*. The diagonal elements (*) are chosen so that each row sums either to one or to zero, depending upon whether the entries are considered probabilities or rates. The parameters π_A, π_G, π_C, and π_T are the equilibrium frequencies of each nucleotide, so that $\pi_A+\pi_G+\pi_C+\pi_T=1$. *R* and *Y* designate purines and pyrimidines, respectively, so $\pi_R=\pi_A+\pi_G$ and $\pi_Y=\pi_C+\pi_T$. The *TI*:*TV* rate ratio for any model is given by:

$$\frac{\pi_A\lambda_{AG}+\pi_G\lambda_{GA}+\pi_C\lambda_{CT}+\pi_T\lambda_{TC}}{\pi_A\lambda_{AY}+\pi_G\lambda_{GY}+\pi_C\lambda_{CR}+\pi_T\lambda_{TR}}$$

where $\lambda_{AY}=\lambda_{AC}+\lambda_{AT}$, and so on. For Hasegawa *et al.*'s model this simplifies to:

$$\frac{(\pi_A\pi_G+\pi_C\pi_T)\alpha}{\pi_R\pi_Y\beta}$$

Thus, when $\pi_A=\pi_G=\pi_C=\pi_T=1/4$, this model reduces to Kimura's, with its *TI*:*TV* rate ratio of α:(2β). Felsenstein's model has *TI*:*TV* rate ratio:

$$\frac{\pi_A\pi_G+\pi_C\pi_T+\left(\frac{\pi_A\pi_G}{\pi_R}+\frac{\pi_C\pi_T}{\pi_Y}\right)\kappa}{\pi_R\pi_Y}$$

which reduces to $\kappa+1/2$ when the base frequencies are all 1/4. While more complicated models are possible and some models proposed do not explicitly include transition bias, both Hasegawa *et al.*'s and Felsenstein's models have been shown to provide a very good fit to observed sequence data, especially when rate variation among sites is taken into account[25,35].

Solutions in new methods of estimation

The problem of saturation results from using the relative numbers of transition and transversion differences observed (pairwise) or inferred (parsimony) among a sample of sequences as a guide to the relative rates of transition and transversion. This problem disappears if multiple changes are dealt with explicitly in a method or if observed differences are corrected for multiple substitutions. If *P* and *Q* are the observed proportions of transitions and transversions between two sequences, Kimura suggested using:

$$\widehat{4\alpha t} = -\log_e(1-2P-Q)+\frac{1}{2}\log_e(1-2Q) \qquad (3)$$

and

$$\widehat{8\beta t} = -\log_e(1-2Q) \qquad (4)$$

to estimate the total numbers of transitions, $4\alpha t$, and transversions, $8\beta t$, that have occurred. The ratio of these is an estimate of the *TI*:*TV* rate ratio that does not suffer from the problem of saturation[26]. This estimator and its analogs for a number of models, both with and without rate variation among sites, can be calculated in the computer package MEGA where they are called *R* (Ref. 27).

The problem that the expectation of the ratio of two random variables is not the same as the ratio of the expectations is still an issue for *R*. In fact, the problem is complicated by the fact that *R* is a ratio of two complicated functions of the numbers of transitions and transversions, not just of the numbers themselves. For Kimura's model, Pollock and Goldstein[28] recently suggested a new estimator, R_{VAR}, that corrects for some of the bias resulting from taking a ratio of random variables and is a weighted average of values for all pairs of sequences in a sample. While R_{VAR} may still be biased, it is accurate in the one case examined[28].

Yang and Kumar[29] have recently proposed a method of correcting tree-based parsimony counts of changes for multiple substitutions. First, a symmetric matrix of the numbers of each type of nucleotide change is compiled using parsimony reconstructions of ancestral bases. The underlying rate matrix and the total time of the tree are recovered using the general reversible model analog of eqns (3) and (4). If Kimura's two-parameter model is assumed, Yang and Kumar's method amounts to counting the numbers of transitions and transversions necessitated by the tree, then correcting them using eqns (3) and (4). For two data sets examined, this corrected parsimony method gives estimates of the pattern of nucleotide substitution similar to those obtained from maximum-likelihood analysis[29].

All three of the problems discussed above, including that of rate variation among sites, can be circumvented by using maximum likelihood to estimate the pattern of nucleotide substitution. This is because maximum likelihood allows for multiple substitutions at a site, can be formulated so that transition bias is a parameter that is estimated directly, and is easily generalized. Yang[30] extended the approach of Felsenstein[31] to include a gamma distribution of substitution rates among sites. This method allows the simultaneous estimation of the substitution matrix, the distribution of rates among sites, and the tree relating the sequences, but has the disadvantage of being quite slow computationally.

Yang[32] subsequently introduced a method in which the gamma distribution is approximated using several categories of rates or a fixed number of rates is assumed, greatly reducing the computational requirements. However, the computational burden of Yang's maximum likelihood is many times greater than either the corrected pairwise method of Pollock and Goldstein[28] or the corrected parsimony method of Yang and Kumar[29]. For large data sets, these faster methods will be the only practical option.

Some maximum-likelihood estimates of the *TI*:*TV* rate ratio for relatively small data sets are given in Box 2, where they are juxtaposed to pairwise or tree-based parsimony estimates for the same kinds of sequences. For most of these data, the classical and best estimates are very similar. However, the direction of differences is impossible to predict, perhaps owing to differences in the relative magnitudes of the problems of estimating transition bias discussed above. In addition, classical and new methods cannot, in general, be expected to give similar estimates[24,33]. Box 2 also illustrates the fact that different types of DNA sequences have strikingly different levels of transition bias.

The future

The dynamics of substitution among the four nucleotides are complex. It is now clear that transition bias is just

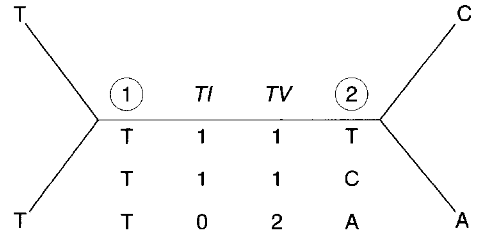
(a)

Sequence I	A	C	A	T	G	C
Sequence II	A	T	A	T	G	C
Sequence III	G	G	G	C	T	C
Sequence IV	A	A	G	A	T	T

(b)

	I	II	III	IV
I	*	1:0	3:2	2:3
II	1	*	3:2	2:3
III	5	5	*	3:1
IV	5	5	4	*

(c)



(d)

Site	*TI*	*TV*	No. MPR
1	1	0	1
2	1.5	1.5	8
3	1	0	1
4	0.67	1.33	3
5	0	1	1
6	1	0	1

Fig. 2. A hypothetical data set, consisting of six variable sites, is shown in (a). In the pairwise method, pairs of sequences are directly compared and the differences between them counted. For example, sequence I and sequence II differ only at site 2 and the difference is a C ↔ T transition. Counts for all pairwise comparisons are given in (b); above the diagonal are the resulting *TI*:*TV* rate ratio estimates for each pair and below the diagonal are the total numbers of differences. The estimates range from infinity for pair I–II to 0.67 for pairs I–IV and II–IV. There is no clear rationale for choosing among these, nor for taking the average, since their expectations may be different. In the tree-based parsimony method, an algorithm like that of Fitch[23] is applied to each site to give the minimum number of changes required by the tree. In (c), sequences I and II share a common ancestor in node 1 and sequences III and IV share a common ancestor in node 2. The three equally parsimonious reconstructions of ancestral bases for site 4 are shown. Each of these requires two changes, but they differ in the numbers of transitions and transversions. The numbers of transition and transversion changes inferred at each site are given in (d). For each site, the number of most-parsimonious reconstructions of ancestral states (MPR) is shown in the right-hand column, and when more than one of these is possible, the numbers of transitions and transversions are averaged over all the possibilities. For example, for site 4, the number of transitions is (1 + 1 + 0)/3 = 0.67 and the number of transversions is (1 + 1 + 2)/3 = 1.33. The total numbers of transitions and transversions are 5.17 and 3.83, respectively, so the resulting *TI*:*TV* rate ratio estimate for the parsimony method is 1.35:1.

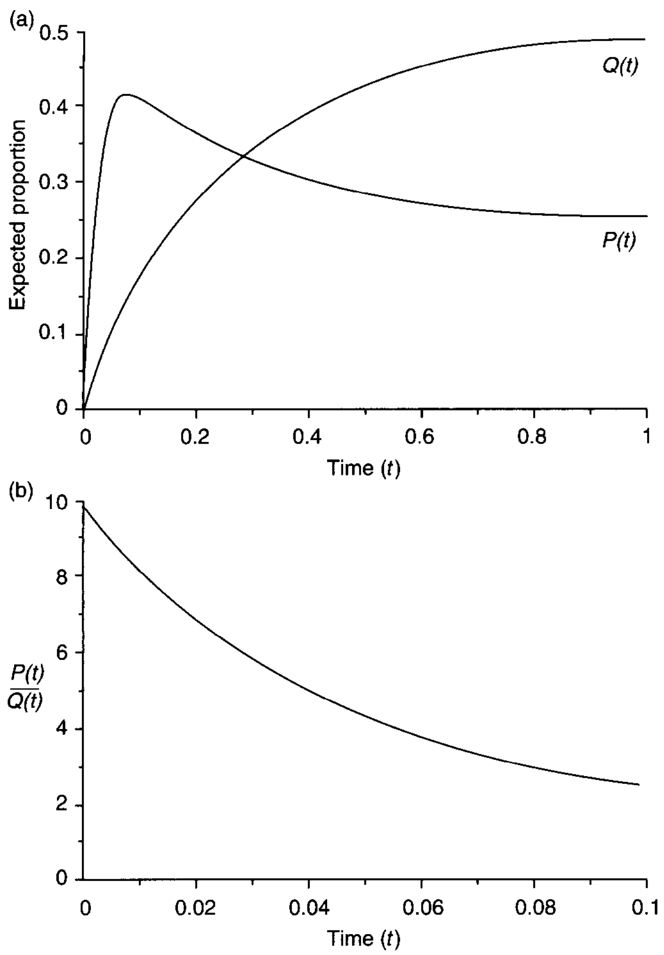


Fig. 3. (a) A plot of the expected proportions of transition and transversion differences between two sequences, given by eqns (1) and (2). The *TI*:*TV* rate ratio, α:(2β), is assumed to be ten. If transversions were disallowed, the proportion of transitions counted between two sequences would eventually approach 1/2. In this example, although transversions can occur, *P*(*t*) rises quickly above the equilibrium value of 1/4 (i.e. towards 1/2) because there is strong bias. After that, as sites already showing transitions experience transversions, the proportion decreases towards 1/4. The time scale of the graph is arbitrary; if $t = 1.0$ is taken to be ten million years, then $\alpha = 1.0 \times 10^{-6}$. At the far right in (a), there has been an average of 44 changes per site. The ratio *P*(*t*):*Q*(*t*), shown in (b), starts out close to ten when *t* is small, then decreases steadily. Notice that the horizontal axis in (b) spans only the first one-tenth of that given in (a). Although it takes a very long time for *P*(*t*) and *Q*(*t*) to reach their equilibrium values, *P*(*t*):*Q*(*t*) falls rapidly from α:(2β). This decrease has been observed consistently in DNA-sequence data.

one of several important factors; uneven base composition and rate variation among sites also appear to be prevalent. In fact, each of the twelve possible base substitutions may happen at a unique rate. Nevertheless, the excess of transitions over transversions remains a striking and consistently observed feature of DNA-sequence change. Thus, the accurate estimation of transition bias is important to our understanding of evolution. The new methods reviewed above are available in program form; interested readers should consult the papers cited here.

It is now possible to make reliable estimates of transition bias, which will improve evolutionary-distance corrections, and, in turn, improve estimates of divergence times and of phylogenetic relationships. In addition, a more detailed knowledge of transition bias should generate interest in determining the relative importance of its various causes. The future may see the development of a framework in which mutational biases and population-level processes, such as selection, can be addressed together.

Box 2. Estimates of transition bias

Type of sequence	Classical estimate	Best available
mtDNA	9.0	10.6
12S rRNA	1.75	6.04
α- and β-globins	0.66	0.62
Pseudo η-globin	2.70	2.55

The classical (pairwise or tree-based parsimony) estimates are from the following sources: for mitochondrial DNA (mtDNA), Brown *et al.*[8] used a modified pairwise method to estimate 90% of all changes to be transitions; for 12S RNA, Sullivan *et al.*[33] used the tree-based parsimony method; for α- and β-globins, Gojobori *et al.*[3] presented parsimony-based estimates for each type of nucleotide change in their Table 2; and for pseudo η-globin, Miyamoto *et al.*[36] used the same method as Gojobori *et al.* to estimate that 73% of all changes are transitions.

The best available estimates are from the following sources: for mtDNA and α- and β-globins, Yang[32] simultaneously inferred the distribution of rates among sites and the parameters of Felsenstein's model; for 12S RNA, Sullivan *et al.*[33] used Yang's maximum-likelihood method; and for pseudo η-globin, Yang *et al.*[25] used maximum likelihood without rate variation among sites to estimate the parameters of Hasegawa *et al.*'s model.

For the α- and β-globin coding sequences, only first and second codon positions were analyzed. These estimates for α- and β-globins are not strictly comparable because they were made from data sets that only partially overlap in terms of species. For all others, classical and best estimates were made from the same data.

Acknowledgements

I am greatly indebted to Chris Simon, Ziheng Yang, Sudhir Kumar, Yasuo Ina, David Pollock, Ross Crozier, Jack Sullivan, Mary-Ellen Ruvolo, Wesley Brown and Tom Kocher for helpful discussions and/or comments on the manuscript. This work was supported by grant P94168 from the Japan Society for the Promotion of Science and was performed in the laboratory of Tomoko Ohta at the National Institute of Genetics in Mishima, Japan.

References

1 Vogel, F. and Rörhborn, G. (1966) **Amino-acid substitutions in haemoglobins and the mutation process,** *Nature* 210, 116–117

2 Fitch, W.M. (1967) **Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations,** *J. Mol. Biol.* 26, 499–507

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

4 Li, W-H., Wu, C-I. and Luo, C-C. (1984) **Nonrandomness of point mutations reflected in nucleotide substitutions in pseudogenes and its evolutionary implications,** *J. Mol. Evol.* 21, 58–71

5 Sankoff, D., Morel, C. and Cedergren, R.J. (1973) **Evolution of 5S RNA and the non-randomness of base replacement,** *Nat. New Biol.* 245, 232–234

6 Hixon, J.E. and Brown, W.M. (1986) **A comparison of small ribosomal RNA genes from the mitochondrial DNA of great apes and humans: sequence, structure, evolution and phylogenetic implications,** *Mol. Biol. Evol.* 3, 1–18

7 Vigilant, L. *et al.* (1991) **African populations and the evolution of human mitochondrial DNA,** *Science* 253, 1503–1507

8 Brown, W.M. *et al.* (1982) **Mitochondrial DNA sequences of primates: the tempo and mode of evolution,** *J. Mol. Evol.* 18, 225–239

9 Curtis, S.E. and Clegg, M.T. (1984) **Molecular evolution of chloroplast DNA sequences,** *Mol. Biol. Evol.* 1, 291–301

10 Goldstein, D.B. and Pollock, D.D. (1994) **Least squares estimation of molecular distance: noise abatement in phylogenetic reconstruction,** *Theor. Popul. Biol.* 45, 219–226

11 Ruvolo, M. *et al.* (1993) **Mitochondrial COII sequences and modern human origins,** *Mol. Biol. Evol.* 10, 1115–1135

12 Jukes, T.H. and Cantor, C.R. (1969) **Evolution of protein molecules,** in *Mammalian Protein Metabolism* (Munro, H.R., ed.), pp. 21–132, Academic Press

13 Kimura, M. (1980) **A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences,** *J. Mol. Evol.* 16, 111-120

14 Watson, J.D. and Crick, F.H.C. (1953) **Genetical implications of the structure of deoxyribonucleic acid,** *Nature* 171, 964-967

15 Topal, M.D. and Fresco, J.R. (1976) **Complementary base pairing and the origin of substitution mutations,** *Nature* 263, 285–289

16 Echols, H. and Goodman, M.F. (1991) **Fidelity mechanisms in DNA replication,** *Annu. Rev. Biochem.* 60, 477–511

17 Drake, J.W. (1991) **Spontaneous mutation,** *Annu. Rev. Genet.* 25, 125–146

18 Crozier, R.H. and Crozier, Y.C. (1993) **The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization,** *Genetics* 133, 97–117

19 Vogel, F. and Kopun, M. (1977) **Higher frequencies of transitions among point mutations,** *J. Mol. Evol.* 9, 159–180

20 Naylor, G.J.P., Collins, T.M. and Brown, W.M. (1995) **Hydrophobicity and phylogeny,** *Nature* 373, 565–566

21 Gutell, R.R., Larson, N. and Woese, C.R. (1994) **Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective,** *Microbiol. Rev.* 58, 10–26

22 Vawter, L. and Brown, W.M. (1993) **Rates and patterns of base change in the small subunit ribosomal RNA gene,** *Genetics* 134, 597–608

23 Fitch, W.M. (1971) **Toward defining the course of evolution: minimum change for a specific tree topology,** *Syst. Zool.* 20, 406–416

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

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

26 Jukes, T.H. (1987) **Transitions, transversions and the molecular clock,** *J. Mol. Evol.* 26, 87–98

27 Kumar, S., Tamura, K. and Nei, M. (1993) **MEGA: Molecular Evolutionary Genetic Analysis, version 1.0,** Pennsylvania State University, USA

28 Pollock, D.D. and Goldstein, D.B. (1995) **A comparison of two methods for reconstructing evolutionary distances from a weighted contribution of transition and transversion differences,** *Mol. Biol. Evol.* 12, 713–717

29 Yang, Z. and Kumar, S. **Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites,** *Mol. Biol. Evol.* (in press)

30 Yang, Z. (1993) **Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites,** *Mol. Biol. Evol.* 10, 1396–1401

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

32 Yang, Z. (1994) **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods,** *J. Mol. Evol.* 39, 306–314

33 Sullivan, J., Holsinger, K.E. and Simon, C. **The effect of topology on estimates of among-site rate variation,** *J. Mol. Evol.* (in press)

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

35 Yang, Z. (1994) **Estimating the pattern of nucleotide substitution,** *J. Mol. Evol.* 39, 105–111

36 Miyamoto, M.M., Slighton, J.L. and Goodman, M. (1987) **Phylogenetic relations of humans and African apes from DNA νη-globin region,** *Science* 238, 369–373

What we don't know about great ape variation

Akiko Uchida

The patterning of intraspecific variation among the great apes is proving more complex than has been recognized previously. The great ape species, as currently defined, may include markedly different subspecies, alternatively, the majority of intraspecific variation may be observed at the populational level within a single subspecies. Recent studies have raised a number of questions about great ape evolutionary biology. How many species of living great apes exist? What was the original dietary adaptation of gorillas? How should we define male orang-utan adulthood?

Akiko Uchida is at the Primate Research Institute, Kyoto University, Inuyama, Aichi 484, Japan.

Studies of the living great apes have focused increasingly on population diversity and have been finding marked intraspecific variation[1-10]. As more individuals representing wider geographic ranges are analyzed, the known range of intraspecific variability may increase. One of the striking aspects of the recent findings is that each great ape species appears to show a distinctive pattern of biological variation in genetic, ecological and morphological features, and in relationships between the three (A. Uchida, PhD Thesis, Harvard University, 1992) (Fig. 1). Understanding the mechanisms of the observed morphological variation, such as ecological adaptation and genetic drift, is particularly important for inferring the fossil hominoid paleobiology.

Traditional classification and genetic variability

Four species of great apes are currently recognized: *Gorilla gorilla* with three subspecies, *Pongo pygmaeus* with two subspecies, *Pan troglodytes* with three subspecies, and *Pan paniscus* (see Fig. 2 for recent geographical distribution). The traditional classification of living primates, like that of many other mammals, is based primarily on gross morphological features and modern geographical distributions. However, it may not exactly reflect relationships inferred from genetic studies, nor the definition of biological species: reproductive isolation[11]. A good example of this is the classification of baboons, once perceived as at least four separate species, currently viewed as only one species[12]. The marked genetic variation found within great apes has raised questions about the traditional species and subspecies classification.

Recent mitochondrial DNA (mtDNA) sequence data (D-loop[2] and COII gene[7]) show that the differences between western lowland gorillas (*Gorilla g. gorilla*) and the two eastern subspecies, eastern lowland gorillas and mountain gorillas (*G. g. graueri* and *G. g. beringei*, respectively) are slightly greater than the differences between the two species of chimpanzees (*Pan troglodytes* and *P. paniscus*)[2,7]. Differences between the two

PII: S0169-5347(96)10007-0