Unequal Base Frequencies and the Estimation of Substitution Rates

Nicole T. Perna* and Thomas D. Kocher†

*Program in Genetics, †Department of Ecology, University of New Hampshire

Recently, Tamura and Nei (1993) presented a new method for estimating the number of nucleotide substitutions between two sequences. Their model of substitution allows different rates of pair-wise transition, pyrimidine transition, and transversion. Heterogeneity of substitution among sites is accommodated by allowing the rate to vary according to a gamma distribution. The model assumes a stationary process (i.e., the observed base composition reflects the nucleotide frequencies at equilibrium). In order to maintain the equilibrium composition, substitutions are weighted by the frequency of the mutant base. The motivation for this weighting arises from an analysis of the patterns and relative rates of base substitutions inferred by parsimony from a distance-based tree. We wish to discuss this analysis of substitution patterns, as well as the sensitivity of divergence estimates to the equilibrium nucleotide frequencies which are assumed.

Table 1 shows the substitutions inferred by Tamura and Nei from 93 human control region sequences. Surprisingly, this matrix suggests that the base composition of the control region is changing over time. The number of G's lost by substitution to another nucleotide is 41.5, while the number of G's created by mutation is 68.5. This suggests a net gain of 27 G's over this phylogeny. A change in base composition seems unlikely for two reasons.

First, nucleotide composition is conserved among hominoid mtDNA sequences over a time period considerably greater than that represented by this data (Kondo et al. 1993). Second, the pattern of mtDNA nucleotide composition in primates, and indeed throughout the animal kingdom, is characterized by a low frequency of G on this strand. If the substitution matrix inferred by parsimony is correct, then the base composition of these sequences is not at equilibrium. Evolving according to the inferred matrix, the composition of these sequences would eventually come to equilibrium at 0.26, 0.25, 0.32, and 0.18 for A, T, C, and G respectively. The observed composition is 0.32, 0.23, 0.31, and 0.13. It seems unlikely that the composition of human mitochondrial genomes is becoming more even than that of their ancestors, by a mechanism which reverses the directional pressure against G found in all known animal mtDNA.

The inferred compositional shift is probably an artifact created by the assumptions used in reconstructing ancestral states. Figure 1a shows the pattern of substitution inferred for three terminal character states, using the method which produced the Tamura and Nei matrix. If we assume that the composition of the variable sites is equal to the control region at a whole, then the probability of transition G → A must be 2.4 times higher than the probability of transition from A → G, in order to maintain this equilibrium composition (fig. 1b). Under these conditions, the alternative character state reconstruction shown in figure 1c is likely to represent the true pattern of substitution at some sites. By failing to account for unequal frequencies of nucleotides, and hence unequal rates of substitution, simplistic applications of parsimony may incorrectly reconstruct ancestral character states, causing an underestimation of the total number of substitutions.

We have observed the same phenomenon in several other published studies (Palumbi and Kessing 1991; Tamura 1992; Knight and Mindell 1993). Note that substitutions can be missed even when comparing closely related sequences, such as humans (Tamura and Nei 1993) or members of the Drosophila nasuta subgroup (Tamura 1992). Phylogenetic analyses based on parsimony as an optimality criterion are quite common in molecular systematics. Failure to account for multiple substitutions can cause parsimony to be inconsistent (i.e., fail to converge to the correct tree as nucleotide sampled increase to infinity). Nonlinear transformations to adjust for multiple hits in parsimony analyses have been described for some models of substitution (Steel et al. 1993). Unfortunately, such transformations have not been described for the models of substitution which are best suited to analysis of mtDNA.

The probabilities of forward and backward transition substitution (fig. 1b) may be even more unequal than Tamura and Nei estimated. In applying their model, they used the average frequency of each nucleo-
tide at all sites in the control region. Base composition varies, however, depending on the degree of selective constraint on a site. Fourfold degenerate third-codon positions are among the least constrained positions of the mitochondrial genome, and base frequencies at these sites are much more unequal than those found in the control region taken as a whole. Selective forces acting on the control region (Kocher and Wilson 1991) constrain base frequencies at some sites. The heterogeneity of selective constraints will be reflected as differences in equilibrium nucleotide composition among rapidly and slowly evolving sites.

We suggest that using the average composition of the control region as an estimate of the equilibrium nucleotide frequency causes an underestimation of the substitution inequalities actually present. If the composition of the most variable sites is similar to that of fourfold degenerate sites (where frequency $A = 0.402$ and $G = 0.054$), the actual base-specific rate of $G \rightarrow A$ substitution may be 7.4-fold greater than the rate of $A \rightarrow G$ substitutions. Since the parsimony method used by Tamura and Nei may have greatly underestimated the number of purine transitions, their conclusion that the pyrimidine transition rate is higher should be reexamined.

While it is not clear that either the average composition of the control region or fourfold degenerate sites truly represent the equilibrium frequency of nucleotides at the variable sites in the control region, it is interesting to compare the results arising from different assumptions. We have reanalyzed the complete human and chimp control region sequences using Tamura and Nei's (1993) model of substitution. We compare the results obtained using the overall composition of the control region, with those obtained using base frequencies at fourfold degenerate sites. The frequencies of $A$, $T$, $C$, and $G$ at the fourfold sites are 0.402, 0.132, 0.411, and 0.054, respectively. The average estimate of $d$ among humans is similar using both control region (0.024 ± 0.006) and fourfold site (0.030 ± 0.009) compositions. However, the choice of equilibrium base composition has a much greater effect on the average $d$ for the more divergent chimpanzee sequences used to estimate the modal substitution rate. If we assume a divergence time of 5 Myr between humans and chimps, we calculate a modal divergence rate of $2.13 \times 10^{-9}$ using the fourfold site composition. This is much higher than the rate calculated from the average D-loop composition (7.5 $\times 10^{-9}$) and would force a revision of the estimated age of the common human ancestor from 160,000 yr to just 71,000 yr.

We expect that the stationary model of substitution developed by Tamura and Nei, when properly applied, will be one of the best methods to estimate divergence for sequences in which the four nucleotide frequencies are unequal. Accounting for unequal nucleotide composition in substitution models is not a trivial matter pertaining only to a few data sets. Most mitochondrial, procarytic, and nuclear genomes exhibit some compositional inequalities. Analyses of ribosomal DNA, in which selective constraints and base composition vary among sites, are likely to present problems similar to those encountered in the analysis of the mitochondrial control region. Studies employing the principle of parsimony to infer patterns of substitution need to address the effect of systematic reconstruction biases among even closely related sequences when nucleotide frequencies are unequal. Unequal base frequencies are an important complication to the study of substitution patterns and rates.

**Table 1**

<table>
<thead>
<tr>
<th>MUTANT</th>
<th>ORIGINAL NUCLEOTIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T  C  A  G</td>
</tr>
<tr>
<td>T</td>
<td>115  2  2  2</td>
</tr>
<tr>
<td>G</td>
<td>112  5  2  2</td>
</tr>
<tr>
<td>A</td>
<td>1   3  4.5</td>
</tr>
<tr>
<td>G</td>
<td>1   3  4.5</td>
</tr>
</tbody>
</table>

**Fig. 1**—Inferring directed matrices of substitution from observed character states using parsimony criteria. The simplest parsimony reconstruction (a) assumes an equal probability of $A \rightarrow G$ and $G \rightarrow A$ substitutions. If the frequency of $A$ and $G$ are unequal, the probability of substitution will be strongly asymmetric (b). In this case, the alternative reconstruction (c) is likely at some sites.
Acknowledgments

We thank M. Hasegawa, K. Thomas, D. Penny, and two anonymous reviewers for thoughtful comments. This work was supported in part by National Science Foundation grant BSR-9007015 to T.D.K.

LITERATURE CITED


SIMON Easteal, reviewing editor

Received May 27, 1994
Accepted October 28, 1994