Unbiased Estimates of the Number of Nucleotide Substitutions When Substitution Rate Varies Among Different Sites

Andrey Rzhetsky, Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA

Received: 20 April 1993 / Revised: 17 August 1993

Abstract. When the number of nucleotides examined is relatively small, the estimators of nucleotide substitutions between DNA sequences often introduce systematic error even if the data used fit the mathematical model underlying the estimation formula. The systematic error of this kind is especially large for models that allow variation in substitution rate among different sites. In the present paper we present a number of formulas that produce virtually bias-free estimates of evolutionary distances for these models.

Key words: Jukes and Cantor's model — Kimura's two-parameter model — Tamura and Nei's model — Gamma distances — Bias correction

Introduction

In the study of molecular evolution it is important to have unbiased estimates of the number of nucleotide substitutions between two DNA sequences. Indicating that all estimators involving logarithmic expressions [e.g., Jukes and Cantor's (1969) estimator] are biased, Tajima (1993) showed that unbiased estimates are obtainable by expanding logarithmic terms into Taylor's series.

We have noticed that similar biases exist in the estimators developed on the assumption that the rate of nucleotide substitution varies from site to site following the gamma distribution (Nei and Gojobori 1986; Jin and Nei 1990) and that the extent of bias is generally greater for these estimators than for those with equal rate. In this paper we present unbiased estimators of the number of nucleotide substitutions per site when the rate varies according to the gamma distribution.

Unbiased Estimates and their Variances

Gamma Versions of the Jukes-Cantor Distance and the Poisson Correction Distance for Amino Acid Sequences

Nei and Gojobori (1986) showed that when the substitution rate varies according to the gamma distribution, Jukes and Cantor's (1969) formula for estimating the number of nucleotide substitutions per site \( \hat{d} \) should be modified as follows.

\[
\hat{d} = \frac{3}{4} \hat{\rho} \left[ \left(1 - \frac{4}{3} \hat{\rho} \right)^{-1/\rho} - 1 \right]
\]  

(1)

where \( \hat{d} \) is an estimate of \( d \), \( \hat{\rho} \) is the proportion of different nucleotides between two sequences compared, and \( \rho \) is a parameter determining the shape and scale of a gamma distribution. (See Nei 1991.) Equation (1) has been obtained under the assumption that \( \hat{\rho} \) is equal to the expected proportion \( E(\rho) \) of nucleotide differences.

In practice, however, a finite number of nucleotides are examined, so \( \hat{\rho} \) may deviate from \( E(\rho) \), and this deviation introduces a systematic bias. This bias can be avoided if we expand the term \( 1 - (4/3)\hat{\rho} \)^{-1/\rho} into Tay-
lor’s series and use the statistical property that under binomial (or multinomial) sampling an unbiased estimator of $E(p^t)$ is given by $\frac{1}{n} \sum_{i=1}^{n} (1-p)^{i-1} (p^{n-i})$ where $k$ and $n$ are the number of nucleotide differences and the total number of nucleotides examined, respectively. (See Tajima 1993.) Here we note that $(1-x)^{-m}$ can be expressed as follows:

$$(1-x)^{-m} = 1 + \sum_{i=1}^{\infty} \frac{m(m+1)...(m+i-1)}{i!} x^i$$

We therefore obtain the following unbiased estimator of $d$ for this case:

$$\hat{d} = \sum_{i=1}^{k} \left( \frac{4}{3a} \right)^{i-1} \frac{k^{(i)}}{i!n^{(i)}} \prod_{j=1}^{i} [(j-1)a + 1]$$

If we use the delta method in statistics, the variance of $\hat{d}$ is given by

$$V(\hat{d}) = \left[ a/(a + 4\hat{d}/3) \right]^{2(a+1)} \frac{k(n-k)(n^2-1)}{2(n^2(n-1))}$$

It should be noted here that this variance is always smaller than the variance of $\hat{d}$ obtained with the biased estimator (1). This is true for all the unbiased estimators that will be considered below.

By analogy, the gamma version of the Poisson correction distance between amino acid sequences can be written as

$$\hat{d} = a [(1-\hat{p})^{-1/\alpha} - 1]$$

An unbiased estimator for this case is given by

$$\hat{d} = a \left( \hat{d}_1 - 1 \right)$$

where

$$\hat{d}_1 = 1 + \sum_{i=1}^{k} \frac{k^{(i)}}{i!d^{-1} n^{(i)}} \prod_{j=1}^{i} [(j-1)a + 1]$$

and the variance of $\hat{d}$ is given by

$$V(\hat{d}) = d_1^{2(a+1)} \hat{p}(1-\hat{p})/n(n-1)$$

**Gamma Version of Kimura’s Distance**

Jin and Nei (1990) showed that when the substitution rate varies following the gamma distribution, Kimura’s (1980) estimator of $d$ becomes

$$\hat{d} = \frac{a}{2} \left[ (1-2\hat{P} - \hat{Q})^{-1/\alpha} + (1-2\hat{Q})^{-1/\alpha} \right]$$

where $\hat{P}$ and $\hat{Q}$ are the proportions of transitional and transversional nucleotide differences, respectively. This again gives a biased estimator. Expanding $(1-x)^{-m}$ into Taylor’s series and estimating $E(Q^t)$ by $\hat{v}^{(i)}/n^{(i)}$ and $E(P)/E(Q)^{2-d}$ by $d^{(i)}d^{(i-1)}n^{(i)}$, where $s$ and $v$ denote the numbers of transitional and transversional nucleotide differences, respectively, we obtain the following unbiased estimator of $d$.

$$\hat{d} = a (d_1/2 + d_2 - 3/2)/2$$

where $d_1$ and $d_2$ are

$$d_1 = 1 + \sum_{i=1}^{s} \left( \frac{2}{a} \right)^{i} \frac{s^{(i)}}{i!n^{(i)}} \prod_{j=1}^{i} [(j-1)a + 1]$$

$$d_2 = 1 + \sum_{i=1}^{s} \left( \frac{2}{a} \right)^{i} \frac{s^{(i)}}{i!n^{(i)}} \prod_{j=1}^{i} [(j-1)a + 1] \times \frac{\prod_{l=0}^{M_1} 2^{l-1} d_1^{(i-1)l}}{l!(l-1)!} \times \frac{\prod_{l=0}^{M_0} 2^{l-1} d_2^{(i-1)l}}{l!(l-1)!}$$

The variance of $\hat{d}$ in equation (9) is given by

$$V(\hat{d}) = \left[ \alpha_1 \hat{p}(1-\hat{p}) - \alpha_1 \alpha_2 \hat{P} \hat{Q} \right] + \alpha_1 \alpha_2 \hat{Q}(1-\hat{Q})/(n-1)$$

where $M_0 = \max(0, i - v)$, $M_1 = \min(i, s)$, $\alpha_1 = d_1^{(i-1)}$, and $\alpha_2 = (d_1^{(i-1)} + d_2^{(i-1)})/2$.

**Tamura and Nei’s Model**

Recently Tamura and Nei (1993) suggested a novel formula for estimating evolutionary distance for the control regions of mitochondrial DNA that are known to have strong transition/transversion and G+C content biases. Their formula for the case of no variation in substitution rate among sites is given by

$$\hat{d} = -\frac{2g_\alpha g_\gamma}{g_\alpha} \log_e \left( 1 - \frac{g_\alpha \hat{P}}{2g_\alpha g_\gamma} - \frac{\hat{Q}}{2g_\gamma} \right)$$

$$-\frac{-2g_\alpha g_\gamma}{g_\alpha} \log_e \left( 1 - \frac{g_\alpha \hat{P}}{2g_\alpha g_\gamma} - \frac{\hat{Q}}{2g_\gamma} \right)$$

$$-2 \left( g_\alpha g_\gamma - \frac{g_\alpha g_\alpha g_\gamma - g_\alpha g_\gamma g_\gamma}{g_\gamma} \right) \times \log_e \left( 1 - \frac{\hat{Q}}{2g_\alpha g_\gamma} \right)$$

where $\hat{P}_1$ is the proportion of transitional differences.
between purines (A and G), \( \hat{P}_2 \) is the proportion of transitional differences between pyrimidines (T and C), \( \hat{Q} \) is the proportion of transversional differences between the sequences under comparison, and \( g_A, g_T, g_C, g_G \), and \( g_r \) are the equilibrium frequencies of nucleotides A, T, C, G, A+G, and T+C, respectively. Using the approach mentioned above, we obtain the following bias-corrected formula:

\[
\hat{d} = \frac{2g_r g_\theta}{g_r^2} d_1 + \frac{2g_r g_\theta}{g_\theta} d_2 + 2 \left( g_r g_\theta - \frac{g_c g_r g_\theta}{g_r^2} - \frac{g_g g_r g_\theta}{g_r^2} \right) d_3
\]

where

\[
d_1 = \sum_{i=1}^{\nu+2} \frac{(i-1)!}{n!} \sum_{j=M_0}^{M_1} \left( \frac{g_r^2}{g_\theta^2} \right)^{j} \left( \frac{g_\theta}{g_r} \right)^{i-j} \frac{j!(i-j)!}{n!(i-j)}
\]

\[
d_2 = \sum_{i=1}^{\nu+2} \frac{(i-1)!}{n!} \sum_{j=M_0}^{M_2} \left( \frac{g_r^2}{g_\theta^2} \right)^{j} \left( \frac{g_\theta}{g_r} \right)^{i-j} \frac{j!(i-j)!}{n!(i-j)}
\]

\[
d_3 = \sum_{i=1}^{\nu+2} \left[ \frac{1}{2g_r g_\theta} \right]^i \frac{1}{n!}
\]

Here \( M_0 = \max(0, -v) \), \( M_1 = \min(s_1, s_2) \), and \( M_2 = \min(s_1, s_2) \). The variance of this \( \hat{d} \) is given by

\[
V(\hat{d}) = \alpha_1^2 V(\hat{P}_1) + \alpha_2^2 V(\hat{P}_2) + \alpha_3^2 V(\hat{Q}) + 2\alpha_1\alpha_2 Cov(\hat{P}_1, \hat{P}_2) + 2\alpha_1\alpha_3 Cov(\hat{P}_1, \hat{Q}) + 2\alpha_2\alpha_3 Cov(\hat{P}_2, \hat{Q})
\]

where \( s_1, s_2, \) and \( \nu \) stand for the numbers of transitional differences between purines, transitional differences between pyrimidines, and transversional differences for the sequences under comparison, respectively. All the variances in expression (18) are computed by \( \frac{V(\hat{d})}{n} = \hat{x}^2 \), whereas the covariances are given by \( Cov(\hat{x}, \hat{y}) = -\hat{x}\hat{y} \), where \( \hat{x} \) and \( \hat{y} \) stand for variables \( \hat{P}_1, \hat{P}_2, \) and \( \hat{Q} \). \( \alpha_1, \alpha_2, \) and \( \alpha_3 \) are given by \( \alpha_1 = \exp(d_1), \alpha_2 = \exp(d_2), \) and

\[
\alpha_3 = \frac{g_r}{g_r^2} \exp(d_1) + \frac{g_\theta}{g_r} \exp(d_2) + \left( 1 - \frac{g_r}{g_\theta} - \frac{g_\theta}{g_r} \right) \exp(d_3)
\]

Table 1. Average values of \( \hat{d} \) obtained by equations (8) and (9) in a computer simulation

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( n = 100 )</th>
<th>( n = 500 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d )</td>
<td>(8)</td>
<td>(9)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.59</td>
<td>0.49</td>
</tr>
<tr>
<td>1.0</td>
<td>1.30</td>
<td>1.00</td>
</tr>
<tr>
<td>1.5</td>
<td>1.99</td>
<td>1.47</td>
</tr>
<tr>
<td>2.0</td>
<td>2.94</td>
<td>2.04</td>
</tr>
</tbody>
</table>

\( a = 0.15 \)

\( a = 0.75 \)

\( a = 2.0 \)

\( n = 100 \)

\( n = 500 \)

\( d \)

\( \hat{P}_2 \), \( \hat{P}_2 \), and \( \hat{Q} \). However, to avoid unnecessary complication of these equations, one can estimate the equilibrium frequencies from all sequences in the data set under analysis and treat them as constant.

Tamura and Nei (1993) also presented the following gamma version of their estimator.

\[
\hat{d} = 2a \left[ \frac{g_r g_\theta}{g_r} \left( 1 - \frac{g_r}{g_\theta} \right) \hat{P}_1 - \frac{\hat{Q}}{2g_r^2} \right]^{-1/\alpha}
\]

\[
+ \frac{g_r g_\theta}{g_r} \left( 1 - \frac{g_r}{g_\theta} \right) \hat{P}_2 - \frac{\hat{Q}}{2g_r^2} \right]^{-1/\alpha}
\]

\[
\left[ \frac{g_r g_\theta}{g_r} \left( 1 - \frac{g_r}{g_\theta} - \frac{g_\theta}{g_r} \right) \right] \left( 1 - \frac{\hat{Q}}{2g_\theta^2} \right)^{-1/\alpha}
\]

An unbiased estimator for this case is given by

\[
\hat{d} = 2a \left[ \frac{g_r g_\theta}{g_r} \left( 1 - \frac{g_r}{g_\theta} \right) \hat{P}_1 - \frac{\hat{Q}}{2g_r^2} \right]^{-1/\alpha}
\]

\[
+ \frac{g_r g_\theta}{g_r} \left( 1 - \frac{g_r}{g_\theta} \right) \hat{P}_2 - \frac{\hat{Q}}{2g_r^2} \right]^{-1/\alpha}
\]

\[
\left[ \frac{g_r g_\theta}{g_r} \left( 1 - \frac{g_r}{g_\theta} - \frac{g_\theta}{g_r} \right) \right] \left( 1 - \frac{\hat{Q}}{2g_\theta^2} \right)^{-1/\alpha}
\]

\[
- \left[ \frac{g_r g_\theta}{g_r} \left( 1 - \frac{g_r}{g_\theta} - \frac{g_\theta}{g_r} \right) \right] \left( 1 - \frac{\hat{Q}}{2g_\theta^2} \right)^{-1/\alpha}
\]
\[ \hat{d} = 2a \left[ \frac{g_1 g_{G}}{g_y} d_4 + \frac{g_4 g_{T}}{g_y} d_5 \\
+ \left( \frac{g_1 g_y}{g_y} - \frac{g_1 g_{G}}{g_y} - \frac{g_4 g_{T}}{g_y} \right) d_6 - g_y g_o \right] \]

where

\[
d_4 = 1 + \sum_{i=1}^{v+s_1} \left( \frac{g_2 g_y}{g_1 g_y} \right)^{l_i} \left( \frac{g_1^{(i-1)}}{g_y^{(i-1)}} \right) \frac{g_y^{(i-1)}}{l_i(l-i)!} \]

\[
d_5 = 1 + \sum_{i=1}^{v+s_2} \left( \frac{g_2 g_y}{g_1 g_y} \right)^{l_i} \left( \frac{g_1^{(i-1)}}{g_y^{(i-1)}} \right) \frac{g_y^{(i-1)}}{l_i(l-i)!} \]

\[
d_6 = 1 + \sum_{i=1}^{v} \left( 1 - \frac{g_2 g_y}{2g_y g_y} \right)^{l_i} \frac{g_y^{(i-1)}}{l_i(n(i))} \sum_{j=1}^{i} [(j-1)a + 1] \]

\[
M_0 = \max[0, i - v], M_1 = \min[i, i], M_2 = \min[i, s_2], \text{and all other notations are as defined above.} \]

The variance of \( \hat{d} \) in equation (21) can be estimated by equation (18) if we redefine \( \alpha_1, \alpha_2, \) and \( \alpha_3 \) as follows:

\[
\alpha_1 = d_4^{a+1}, \alpha_2 = d_5^{a+1}, \text{and} \]

\[
\alpha_3 = \frac{g_1 g_y}{g_2 g_y} d_4^{a+1} + \frac{g_1 g_y}{g_2 g_y} d_5^{a+1} + \left( 1 - \frac{g_2 g_y}{g_2 g_y} \right) d_3^{a+1} \]

Discussion

The unbiased estimators suggested above are obviously more complicated than the “biased” estimators. However, the extent of bias is quite high when the number of nucleotides examined is relatively small or when the two sequences compared are highly divergent. This can be seen from Table 1, where the average values of \( \hat{d} \) estimated by equations (8) and (9) are presented. These results were obtained by conducting computer simulation of DNA sequence divergence and applying equations (8) and (9) for the same set of data. Table 1 shows that the extent of bias is greater when the number of nucleotides \( (n) \) is small and the \( a \) value is small than when these are large. Particularly, when \( d = 2.0, n = 100, \) and \( a = 0.15, \) equation (8) gives nearly a 50% overestimate. Of course, \( a = 0.15 \) is rather extreme value, but a similar value was observed for the control region of mtDNA when this entire region was considered (Tamura and Nei 1993). However, there is no need to use the formulas presented here when \( n \) is large (say, \( n > 1,000 \)) with a relatively large value of \( a. \)

The matter of reducing estimation bias is not as trivial as it may seem to be. The problem is that it is not always easy to predict a priori whether the systematic errors in the estimates of evolutionary distances will affect the results of data analysis. Sometimes these errors may cause serious misinterpretation of the results of phylogenetic inference. For example, Zharkikh and Li (1992) performed a computer simulation of the evolution of nucleotide sequences following Jukes and Cantor’s (1969) substitution scheme to test the consistency of the neighbor-joining tree-making method (Saitou and Nei 1987). Zharkikh and Li (1992) used relatively short model sequences (500 nucleotides), but the expected evolutionary distance between some sequences was set to be up to two nucleotide substitutions per site. The authors’ conclusion was that the neighbor-joining method “with corrected distances performs poorly when the sequence length is short” (i.e., it chooses an incorrect tree more often than the true one). This statement is misleading, since it is the estimator of evolutionary distance rather than the tree-making method that is responsible for these results. If Tajima’s (1993) formula were used instead of the original Jukes and Cantor’s (1969) estimator, the neighbor-joining method would probably have shown a better performance than their results showed. (Note also that sequence data with \( d \) as large as 2 are rarely used for actual phylogenetic inference because the nucleotide differences are close to the saturation level.) At any rate, it is expected that if we use Tajima’s (1993) formulas or the new correction formulas presented in this paper, we can improve the performance of phylogenetic inference considerably.

Computer Program

A computer program for estimating evolutionary distance based on formulation in the paper is available on request.

Acknowledgments. This work was supported by grants from NIH and NSF to M. Nei.
References
