J Mol Evol (1993) 36:96–99

Journal of
Molecular Evolution


*Letter to the Editor*

# Unbiased Estimation of the Rates of Synonymous and Nonsynonymous Substitution

Wen-Hsiung Li

Center for Demographic and Population Genetics, University of Texas, PO Box 20334, Houston, TX 77225, USA

**Summary.** The current convention in estimating the number of substitutions per synonymous site ($K_S$) and per nonsynonymous site ($K_A$) between two protein-coding genes is to count each twofold degenerate site as one-third synonymous and two-thirds nonsynonymous because one of the three possible changes at such a site is synonymous and the other two are nonsynonymous. This counting rule can considerably overestimate the $K_S$ value because transitional mutations tend to occur more often than transversional mutations and because most transitional mutations at twofold degenerate sites are synonymous. A new method that gives unbiased estimates is proposed. An application of the new and the old method to 14 pairs of mouse and rat genes shows that the new method gives a $K_S$ value very close to the number of substitutions per fourfold degenerate site whereas the old method gives a value 30% higher. Both methods give a $K_A$ value close to the number of substitutions per nondegenerate site.

**Key words:** Transition — Transversion — Synonymous rate — Nonsynonymous rate — Estimation methods

In studying the evolution of protein-coding genes it is useful to distinguish between synonymous and nonsynonymous (amino acid changing) substitutions, and several methods have been proposed for estimating the numbers of the two types of substitution between two sequences (e.g., Miyata and Yasunaga 1980; Li et al. 1985; Nei and Gojobori 1986). The current convention in estimating these two numbers is to count each twofold degenerate site as one-third synonymous and two-thirds nonsynonymous because at such a site one of the three possible changes is synonymous and the other two are nonsynonymous. However, as will be explained below, this counting rule tends to overestimate the synonymous rate and underestimate the nonsynonymous rate. The purpose of this note is to propose a new method that gives unbiased estimates.

The method to be proposed is a modification of Li et al.'s (1985) method, which is as follows. Consider, for example, the nuclear genetic code. First, the nucleotide sites in a sequence are classified into nondegenerate, twofold degenerate, and fourfold degenerate sites. A site is nondegenerate if all possible changes at that site are nonsynonymous, twofold degenerate if one of the three possible changes is synonymous, and fourfold degenerate if all possible changes at the site are synonymous. The only case of a threefold degenerate site is the third position of the three isoleucine codons, which, for simplicity, is treated as a twofold degenerate site. Using the above rules, one first counts the numbers of the three types of sites in each of the two sequences compared and then computes the averages, denoting them by $L_0$ (nondegenerate), $L_2$ (twofold) and $L_4$ (fourfold), respectively.

Second, compare the two sequences codon by codon and infer the nucleotide differences between each pair of codons. Classify each difference according to the type of site at which it has occurred. The nucleotide differences in each class are further classified into transitional ($S_i$) and transversional ($V_i$) differences ($i = 0, 2, 4$). In the class of twofold degenerate sites transitions are synonymous and transversions are nonsynonymous. There are two exceptions: the first position of the arginine codons (CGA, CGG, AGA, and AGG) and the last position in the three isoleucine codons (AUU, AUC, and AUA). In all these exceptional cases, all synonymous changes are included in $S_2$ and all nonsynonymous changes are included in $V_2$. Let $P_i = S_i/L_i$ and $Q_i = V_i/L_i$, which are, respectively, the proportions of transitional differences and transversional differences at $i$-fold degenerate sites between the two sequences.

Third, use Kimura's (1980) two-parameter method to estimate the numbers of transitional ($A_i$) and transversional ($B_i$) substitutions per $i$-th type site. The means are given by

$$A_i = (1/2)\, n\, (a_i) - (1/4)\, n\, (b_i) \quad (1)$$

$$B_i = (1/2)\, n\, (b_i) \quad (2)$$

and the variances are given by

$$V(A_i) = [a_i^2 P_i + c_i^2 Q_i - (a_i P_i + c_i Q_i)^2]/L_i \quad (3)$$

$$V(B_i) = b_i^2 Q_i (1 - Q_i)/L_i \quad (4)$$

where $a_i = 1/(1 - 2P_i - Q_i)$, $b_i = 1/(1 - 2Q_i)$ and $c_i = (a_i - b_i)/2$. The total number of substitutions per $i$-th type site, $K_i$, is given by

$$K_i = A_i + B_i \quad (5)$$

Finally, let $K_S$ be the number of (synonymous) substitutions per synonymous site and $K_A$ the number of (nonsynonymous) substitutions per nonsynonymous site. Follow the convention to count one-third of a twofold degenerate site as synonymous and two-thirds as nonsynonymous and obtain

$$K_S = (L_2 A_2 + L_4 K_4)/(L_2/3 + L_4) \quad (6)$$

$$K_A = (L_2 B_2 + L_0 K_0)/(2L_2/3 + L_0) \quad (7)$$

The variances of $K_S$ and $K_A$ are given in Li et al. (1985).

The above rule of counting one-third of a twofold degenerate site as synonymous and two-thirds as nonsynonymous tends to overestimate $K_S$ (and underestimate $K_A$) because transitional substitutions tend to occur more often than transversional substitutions and because most transitional changes at twofold degenerate sites are synonymous changes. (See data below.)

To overcome the above problem I propose to take the weighted average $(L_2 A_2 + L_4 A_4)/(L_2 + L_4)$ as an estimate of the average transitional rate at twofold and fourfold degenerate sites; $L_2 A_2$ is the total number of transitional substitutions at twofold degenerate sites between the two sequences, $L_4 A_4$ is the corresponding number at fourfold degenerate sites, and $L_2 + L_4$ is the total number of twofold-plus-fourfold degenerate sites. The number of (synonymous) substitution per synonymous site can then be computed by

$$K_S = (L_2 A_2 + L_4 A_4)/(L_2 + L_4) + B_4 \quad (8)$$

Similarly, I propose to take the weighted average $(L_0 B_0 + L_2 B_2)/(L_0 + L_2)$ as an estimate of the average transversional rate at nondegenerate and twofold degenerate sites and compute the number of (nonsynonymous) substitutions per nonsynonymous site by

$$K_A = A_0 + (L_0 B_0 + L_2 B_2)/(L_0 + L_2) \quad (9)$$

Recently, Pamilo and Bianchi (1993) have also proposed to use these formulas. One can show that the variances of $K_S$ and $K_A$ are given by

$$V(K_S) = [L_2^2\, V(A_2) + L_4^2\, V(A_4)]/(L_2 + L_4)^2 + V(B_4) - b_4 Q_4 [2a_4 P_4 - c_4(1 - Q_4)]/(L_2 + L_4) \quad (10)$$

$$V(K_A) = V(A_0) + [L_0^2\, V(B_0) + L_2^2\, V(B_2)]/(L_0 + L_2)^2 - b_0 Q_0 [2a_0 P_0 - c_0(1 - Q_0)]/(L_0 + L_2) \quad (11)$$

where $V(A_i)$ and $V(B_i)$, $i = 0, 2, 4$, are given by Eqs. (3) and (4).

Table 1 shows the $K_S$ and $K_A$ values computed by the new and old methods for 14 pairs of mouse and rat genes. In all cases the $K_S$ value estimated by the old method is higher than that estimated by the new method. When all 14 genes are pooled together, the old method gives a $K_S$ value of 0.183, which is about 30% higher than that (0.144) obtained from the new method.

To see which of the two estimates is more reasonable, one can compare them with the number of substitutions per fourfold degenerate site estimated from the same set of genes because all possible changes at a fourfold degenerate site are synonymous, so it is truly a synonymous site in the sense

**Table 1.** Numbers of substitutions per synonymous site ($K_S$) and per nonsynonymous site ($K_A$) between mouse and rat genes estimated by the old and new methods[a]

| Gene[b] | $K_S$ | | $K_A$ | |
|---|---|---|---|---|
| | Old | New | Old | New |
| Nucleolin | 0.139 | 0.110 | 0.0277 | 0.0281 |
| HSC70α | 0.187 | 0.148 | 0.0007 | 0.0008 |
| TCP-1 | 0.230 | 0.171 | 0.0110 | 0.0130 |
| Chaperonin | 0.098 | 0.077 | 0.0008 | 0.0007 |
| ODC | 0.256 | 0.205 | 0.0123 | 0.0129 |
| β2-ADR | 0.205 | 0.169 | 0.0208 | 0.0227 |
| Go-α | 0.127 | 0.093 | 0.0024 | 0.0025 |
| GAPDH | 0.213 | 0.167 | 0.0105 | 0.0104 |
| ApoB100 | 0.278 | 0.233 | 0.0700 | 0.0757 |
| Prion protein | 0.216 | 0.178 | 0.0134 | 0.0151 |
| HPRT | 0.145 | 0.091 | 0.0098 | 0.0109 |
| αB Crystallin | 0.226 | 0.182 | 0.0000 | 0.0000 |
| GRP78 | 0.278 | 0.239 | 0.0050 | 0.0061 |
| Cyclophilin | 0.113 | 0.078 | 0.0105 | 0.0118 |
| All | 0.183 | 0.144 | 0.0137 | 0.0148 |

[a] For sequence data sources, see O'hUigin and Li (1992)
[b] Gene names: HSC70α, heat-shock cognate protein 70α; TCP-1, T-complex protein 1; ODC, ornithine decarboxylase; α2-ADR, α2 adrenergic receptor; Go-α, G protein α subunit; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; ApoB100, apolipoprotein 100 (the LDL receptor binding region); HPRT, hypoxanthine phosphoribosyl transferase; and GRP78, glucose regulatory protein 78 kD

**Table 2.** Numbers of transitional and transversional substitutions per site at nondegenerate, twofold degenerate, and fourfold degenerate sites of codons between mouse and rat genes[a]

| Type of substitution | Nondegenerate | Twofold degenerate | Fourfold degenerate |
|---|---|---|---|
| Transition | 0.009 ± 0.001 | 0.097 ± 0.006 | 0.084 ± 0.007 |
| Transversion | 0.004 ± 0.001 | 0.011 ± 0.002 | 0.053 ± 0.005 |
| Total | 0.013 ± 0.001 | 0.108 ± 0.007 | 0.137 ± 0.008 |

[a] The numbers are estimated by pooling together all the genes used in Table 1 and "Total" refers to the sum of transitional and transversion substitutions per site

of the word. As can be seen from Tables 1 and 2, the $K_S$ value (0.144) estimated by the new method is very close to the number of substitutions per fourfold degenerate site (0.137) whereas that (0.183) estimated by the old method is considerably larger. The overestimation by the old method arises because transitions occur more frequently than transversions; for example, for the 14 genes used the numbers of transitions and transversions per fourfold degenerate site are 0.084 and 0.053, respectively, although at each site there is only one type of transition but two types of transversion. The new method implicitly assumes that the transitional rate at a twofold degenerate site is the same as that at a fourfold degenerate site. Table 2 shows that the number of transitional substitutions per twofold degenerate site (0.097 ± 0.006) is in fact very close to that at a fourfold degenerate site (0.084 ± 0.007); the two numbers are not statistically different. In conclusion, the new method gives better estimates for the $K_S$ value than does the old method.

It is also interesting to note from Table 1 that the value estimated by the new method is slightly less variable among genes than is the value estimated by the old method. This is because the new method is less dependent on the ratio of the numbers of twofold and fourfold degenerate sites in a gene. However, since the $K_S$ values estimated by the new method still show a two- to threefold variation, the synonymous rate is not uniform among genes.

In the majority of cases for the 14 genes in Table 1 the old method gives a smaller $K_A$ value than does the new method. However, the differences between the two estimates are very small. This is because the number of nondegenerate sites in a gene is usually considerably larger than that of twofold degenerate sites, so the $K_A$ value in a gene is largely determined by the nondegenerate sites. For this reason, the two methods give similar estimates for the $K_A$ value, though the rules of counting twofold degenerate sites are different. Indeed, when all 14 genes are pooled together the $K_A$ values estimated by the new and old methods (0.0148 and 0.0137, respectively) are both very close to the number of substitutions per nondegenerate site (0.013, Table 2).

I have also used both methods to estimate the $K_S$ and $K_A$ values for 45 genes from human and mouse (or rat) and arrived at similar conclusions to those above. In particular, the transitional rate at nondegenerate sites is almost exactly the same as that at twofold degenerate sites, though this is not the case in Table 2.

In comparing the rate of synonymous substitution with the rates of substitution in noncoding regions, e.g., introns or intergenic regions, it is more reasonable to use the $K_S$ value estimated by the new method than that by the old method because the former is closer to the number of substitutions per site at fourfold degenerate sites than is the latter. This is another reason for replacing the old method by the new one.

A computer program for the present method is available from the author upon request; please send a formatted IBM PC-compatible floppy disk or, better, send an E-mail message to GSBS005@UTSPH.BITNET.

*Acknowledgments.* I thank P. Pamilo for sending me a copy of his manuscript and L. Shimmin for comments. This study was supported by NIH grant GM30998.

## References

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

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

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

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

O'hUigin C, Li W-H (1992) The molecular clock ticks regularly in muroid rodents and hamsters. J Mol Evol, in press

Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes. Mol Biol Evol, in press

Received May 20, 1992/Revised and accepted August 12, 1992