Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# Improved variance estimators for one- and two-parameter models of nucleotide substitution

Hsiuying Wang [a], Yun-Huei Tzeng [b], Wen-Hsiung Li [b,c,*]

[a] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
[b] Genomics Research Center, Academia Sinica, Taipei, Taiwan
[c] Department of Evolution and Ecology, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

ARTICLE INFO

*Article history:*
Received 26 March 2008
Received in revised form
23 April 2008
Accepted 28 April 2008
Available online 10 May 2008

*Keywords:*
Substitution model
Variance estimator
Taylor expansion
Empirical formulas

ABSTRACT

The current variance estimators for Jukes and Cantor's one-parameter model and Kimura's two-parameter model tend to underestimate the true variances when the true proportion of differences between the two sequences under study is not small. In this paper, we developed improved variance estimators, using a higher-order Taylor expansion and empirical methods. The new estimators outperform the conventional estimators and provide accurate estimates of the true variances.

## 1. Introduction

A basic process in the evolution of DNA sequences is the substitution of one nucleotide for another during evolution. The substitution of one allele for another in a population generally takes thousands of years or longer to complete, so the process cannot be directly observed. To detect evolutionary changes in a DNA sequence, we need to compare two sequences that have descended from a common ancestral sequence.

If two sequences of length $L$ differ from each other at $X$ sites, the proportion of differences, $X/L$, is referred to as the observed or uncorrected divergence. When the degree of divergence between the two sequences compared is small, the chance for more than one substitution to have occurred at a site is negligible, and the number of observed differences between the two sequences is close to the actual number of substitutions. However, if the degree of divergence is substantial, the observed number of differences is likely to be smaller than the actual number of substitutions due to multiple hits at the same site. Many methods have been proposed to correct for multiple hits (Holmquist, 1971; Jukes and Cantor, 1969; Kaplan and Risko, 1982; Kimura, 1980, 1981; Lanave et al., 1984). The simplest and most frequently used models are Jukes and Cantor's (1969) one-parameter model and Kimura's, (1980) two-parameter model.

Jukes and Cantor's one-parameter model assumes that substitutions occur with equal probability, say $\alpha$, among the four nucleotide types. Since the time of divergence between two sequences is usually unknown, we cannot estimate $\alpha$ directly. Instead, we compute $K$, the number of substitutions per site since the time of divergence between the two sequences. In the one-parameter model case, $K = 2(3\alpha t)$, where $3\alpha t$ is the expected number of substitutions per site in a single lineage. Jukes and Cantor (1969) derived the following formula:

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{p}\right) \tag{1}$$

where $\hat{p} = X/L$ is the observed proportion of different nucleotides between the two sequences. The following approximated estimator for the sampling variance was derived by Kimura and Ohta (1972) and has been commonly used in the literature.

$$V(K) = \frac{\hat{p} - \hat{p}^2}{L(1 - (4/3)\hat{p})^2} \tag{2}$$

In the case of the two-parameter model (Kimura, 1980), the differences between two sequences are classified into transitions and transversions. Let $\hat{P} = X_1/L$ and $\hat{Q} = X_2/L$ be the observed proportions of transitional and transversional differences between the two sequences, respectively, where $X_1$ and $X_2$ are the numbers of transitional and transversional differences between the two

* Corresponding author at: Department of Evolution and Ecology, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA. Tel.: +1 773 702 3104; fax: +1 773 702 9740.
*E-mail address:* whli@uchicago.edu (W.-H. Li).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.04.034

sequences. Then the number of nucleotide substitutions per site between the two sequences, $K_2$, is estimated by

$$K_2 = \frac{1}{2}1\mathrm{n}\left(\frac{1}{1-2\hat{P}-\hat{Q}}\right)+\frac{1}{4}1\mathrm{n}\left(\frac{1}{1-2\hat{Q}}\right) \tag{3}$$

The sampling variance is approximately given by

$$V(K_2)=\frac{1}{L}\left[\hat{P}\left(\frac{1}{1-2\hat{P}-\hat{Q}}\right)^2+\hat{Q}\left(\frac{1}{2-4\hat{P}-2\hat{Q}}+\frac{1}{2-4\hat{Q}}\right)^2 - \left(\frac{\hat{P}}{1-2\hat{P}-\hat{Q}}+\frac{\hat{Q}}{2-4\hat{P}-2\hat{Q}}+\frac{\hat{Q}}{2-4\hat{Q}}\right)^2\right] \tag{4}$$

Since the above two variance estimators underestimate the true variances in most circumstances, we derive improved estimators for estimating the sampling variances, using a higher-order Taylor expansion and empirical methods. Our simulation results show that the new estimators outperform the conventional variance estimators and provide accurate estimates of the sampling variances.

## 2. Methods

Because Eq. (1) involves the log function, it is not easy to directly calculate the variance. So we employ the Taylor expansion to expand the log function at $X = Lp$.

By Taylor expansion at $X = Lp$ to second order, we have

$$\begin{aligned}-\frac{3}{4}1\mathrm{n}\left(1-\frac{4}{3}\frac{X}{L}\right) \approx\ & -\frac{3}{4}1\mathrm{n}\left(1-\frac{4}{3}p\right)\\ &+\left(\frac{X}{L}-p\right)/(1-(4/3)p)\\ &+\left(\frac{X}{L}-p\right)^2 2/(3(1-(4/3)p)^2)\end{aligned} \tag{5}$$

From the formula

$$Var(Y)=E(Y^2)-(EY)^2$$

where $Y$ is a random variable, the variance of $K$ can be expressed as

$$\begin{aligned}Var(K) = E&\left[\left(-\frac{3}{4}1\mathrm{n}\left(1-\frac{4}{3}\frac{X}{L}\right)\right)^2\right]\\ &-\left[E\left(-\frac{3}{4}1\mathrm{n}\left(1-\frac{4}{3}\frac{X}{L}\right)\right)\right]^2\end{aligned} \tag{6}$$

From Eq. (5), the first term in Eq. (6) is

$$\begin{aligned}&E\left[\left(-\frac{3}{4}1\mathrm{n}\left(1-\frac{4}{3}\frac{X}{L}\right)\right)^2\right]\\ &\quad=\frac{9}{16}\ln^2\left(1-\frac{4}{3}p\right)+\frac{p(1-p)}{L}\frac{1}{(1-(4/3)p)^2}\\ &\qquad+\frac{4}{9(1-(4/3)p)^4}E\left(\frac{X}{L}-p\right)^4-\frac{3}{2}\frac{p(1-p)}{L}\ln\left(1-\frac{4}{3}p\right)\\ &\qquad\times\frac{2}{3(1-(4/3)p)^2}+o\left(\frac{1}{L^2}\right)\end{aligned} \tag{7}$$

From Eq. (5), the second term in Eq. (6) is

$$\begin{aligned}&E\left[\left(-\frac{3}{4}1\mathrm{n}\left(1-\frac{4}{3}\frac{X}{L}\right)\right)\right]^2\\ &\quad=\frac{9}{16}1\mathrm{n}^2\left(1-\frac{4}{3}p\right)-\frac{3}{2}\frac{p(1-p)}{L}1\mathrm{n}\left(1-\frac{4}{3}p\right)\\ &\qquad\times\frac{2}{3(1-(4/3)p)^2}+\frac{4}{9(1-(4/3)p)^4}\frac{p^2(1-p^2)}{L^2}+o\left(\frac{1}{L^2}\right)\end{aligned} \tag{8}$$

From Eqs. (7) and (8) and the fact

$$\begin{aligned}E\left(\frac{X}{L}-p\right)^4 &= \frac{p(1-p)(1-6p(1-p)+3np(1-p))}{L^3}\\ &\approx\frac{3p^2(1-p)^2}{L^2}\end{aligned}$$

we have

$$Var\left(-\frac{3}{4}\ln\left(1-\frac{4}{3}\frac{X}{L}\right)\right)\approx\frac{p(1-p)}{L(1-(4/3)p)^2}+\frac{8p^2(1-p)^2}{9L^2(1-(4/3)p)^4} \tag{9}$$

Our simulation study showed that when $p$ is small, the variance estimator (9) provides a better estimator for the true variance than the estimator (2).

Thus, when $p$ is small, we can directly use the estimator (9) as an improved estimator for the variance. However, when $p$ is not small, the estimator (9) is not good enough to approximate the true variance because some higher-order terms become non negligible. Therefore, we use Eq. (9) to propose the following form of a new estimator:

$$a(\hat{p})\frac{\hat{p}(1-\hat{p})}{L(1-(4/3)\hat{p})^2}+b(\hat{p})\frac{8\hat{p}^2(1-\hat{p})^2}{9L^2(1-(4/3)\hat{p})^4} \tag{10}$$

for the one-parameter model, where $a(\hat{p})$ and $b(\hat{p})$ can be derived empirically by simulation, so that the new estimator can approximate the true variance more accurately than formula (9)

For the two-parameter model, we expand the function

$$f(X_1,X_2)=-\frac{1}{2}\ln\left(1-2\frac{X_1}{L}-\frac{X_2}{L}\right)-\frac{1}{4}\ln\left(1-2\frac{X_2}{L}\right)$$

in Eq. (3) at $X_1 = LP$ and $X_2 = LQ$ by using the Taylor expansion to the second order. Then, we have

$$\begin{aligned}f(X_1,X_2)\approx\ & -\frac{1}{2}\ln(1-2P-Q)-\frac{1}{4}\ln(1-2Q)\\ &+(X_1-PL)\frac{1}{L(1-2P-Q)}\\ &+(X_2-QL)\frac{1}{2L}\left(\frac{1}{1-2P-Q}+\frac{1}{1-2Q}\right)\\ &+\frac{1}{2}\left\{(X_1-PL)^2\frac{1}{L^2}\frac{2}{(1-2P-Q)^2}\right.\\ &+2(X_1-PL)(X_2-QL)\frac{1}{L^2(1-2P-Q)^2}\\ &\left.+(X_2-QL)^2\frac{1}{L^2}\left(\frac{1}{2(1-2P-Q)^2}+\frac{1}{(1-2Q)^2}\right)\right\}\end{aligned} \tag{11}$$

From the formula

$$Var(f(X_1-X_2))=E(f^2(X_1,X_2))-(Ef(X_1,X_2))^2$$

and tedious calculations, we obtain

$$\begin{aligned}V(K_2)\approx\frac{1}{L}&\left[P\left(\frac{1}{1-2P-Q}\right)^2\right.\\ &+Q\left(\frac{1}{2-4P-2Q}+\frac{1}{2-4Q}\right)^2\\ &\left.-\left(\frac{P}{1-2P-Q}+\frac{Q}{2-4P-2Q}+\frac{Q}{2-4Q}\right)^2\right]+S\end{aligned} \tag{12}$$

where

$$S=[-16P^4(3-36Q+132Q^2-200Q^3+108Q^4)\\+(-1+Q)^2Q(-12+89Q-272Q^2+424Q^3-336Q^4\\+108Q^5)+32P^3(-3+39Q-168Q^2+332Q^3\\-308Q^4+108Q^5)+8P^2(8-115Q+574Q^2-1402Q^3\\+1820Q^4-1208Q^5+324Q^6)+8P(-2+33Q\\-191Q^2+562Q^3-942Q^4+916Q^5-484Q^6\\+108Q^7)]/(8L^2(1-2Q)^4(-1+2P+Q)^4)$$

By an argument similar to that for the one-parameter model, we propose, on the basis of Eq. (12), the following form of a new estimator

$$c(\hat{P},\hat{Q})\left[\hat{P}\left(\frac{1}{1-2\hat{P}-\hat{Q}}\right)^2+\hat{Q}\left(\frac{1}{2-4\hat{P}-2\hat{Q}}+\frac{1}{2-4\hat{Q}}\right)^2\\-\left(\frac{\hat{P}}{1-2\hat{P}-\hat{Q}}+\frac{\hat{Q}}{2-4\hat{P}-2\hat{Q}}+\frac{\hat{Q}}{2-4\hat{Q}}\right)^2\right]+d(\hat{P},\hat{Q})\hat{S} \qquad (13)$$

for the two-parameter model, where $\hat{S}$ is the estimator of $S$ by replacing $P$ and $Q$ in $\hat{S}$ by $\hat{P}$ and $\hat{Q}$, respectively.

## 3. Results and discussion

From the forms of Eqs. (10) and (13), we employ an empirical method to find suitable $a(\hat{p})$, $b(\hat{p})$, $c(\hat{P},\hat{Q})$ and $d(\hat{P},\hat{Q})$ such that the new estimators can be close to the true variances. There are many options of $a(\hat{p})$, $b(\hat{p})$, $c(\hat{P},\hat{Q})$ and $d(\hat{P},\hat{Q})$ which can lead to better estimators for the variances of the one- and two-parameter models.

To obtain general formulas for $a(\hat{p})$ and $b(\hat{p})$ in the one-parameter model, we use simulation to profile the relation of the true variance and the estimator (9) first, and then adopt the model selection method to derive $a(\hat{p})$ and $b(\hat{p})$. We fix $a(\hat{p})=b(\hat{p})=1$ to obtain the new estimators at first. Because the difference between the true variances and new estimators increases exponentially as $\hat{p}$ increases, we assume that the coefficient terms in Eq. (10) are functions of $\hat{p}$ and use the nonlinear regression method to obtain the approximation formulas of $a(\hat{p})$ and $b(\hat{p})$. Although there are many possible choices of $a(\hat{p})$ and $b(\hat{p})$, we choose those that can perform well under all different sequence length $L$ in our simulation. The derivation of coefficient terms $c(\hat{P},\hat{Q})$ and $d(\hat{P},\hat{Q})$ in Eq. (13) of the two-parameter model is similar to the one-parameter model.

From the above simulations, we propose

$$V^*(K)=0.6e^{9\hat{p}}\frac{\hat{p}(1-\hat{p})}{L(1-(4/3)\hat{p})^2}+\frac{8}{9}\frac{\hat{p}^2(1-\hat{p})^2}{L^2(1-(4/3)\hat{p})^4} \qquad (14)$$

and

$$V^*(K_2)=\frac{0.56e^{10(\hat{P}+\hat{Q})}}{L}\left[\hat{P}\left(\frac{1}{1-2\hat{P}-\hat{Q}}\right)^2\\+\hat{Q}\left(\frac{1}{2-4\hat{P}-2\hat{Q}}+\frac{1}{2-4\hat{Q}}\right)^2\\-\left(\frac{\hat{P}}{1-2\hat{P}-\hat{Q}}+\frac{\hat{Q}}{2-4\hat{P}-2\hat{Q}}+\frac{\hat{Q}}{2-4\hat{Q}}\right)^2\right]+\hat{S} \qquad (15)$$

to be the new estimators of the variances for the one- and two-parameter models, respectively.

To test the performances of formulas (14) and (15), we generate DNA sequences by using the *evolver* program in PAML package (Yang, 1997). Several combinations of parameter values are used to generate different data sets: sequence length ($L=500$, 1000 and 5000) and the expected number of nucleotide substitutions per site (0.1–0.7). For each data set, we generate 1000 pairs of sequences and calculate their corresponding $K$ values from formula (1). Hence, we can calculate the sample variance of these 1000 values of $K$ and use it as the true variance of each data set. A similar simulation procedure is used for Kimura's two-parameter model, and the ratio of transition/transversion is set to be 1, 2 and 5.

Tables 1 and 2 show the comparisons of the new estimators (14) and (15) and the conventional estimators (2) and (4). For the one-parameter model, when the number of substitutions per site is low, the conventional estimators are not far from the true estimators. For example, when the expected number of nucleotide substitutions per site is 0.1, the conventional estimator underestimates the true variance within a tolerable region. However, as the divergence increases, the performance becomes poor. When the divergence is greater than 0.2, the conventional estimators seriously underestimate the true variance, for all the different sequence lengths studied.

As seen from Table 1, the improved estimator can accurately estimate the true variance for the case where the expected number of nucleotide substitutions per site is 0.1 or 0.2. When the expected number of nucleotide substitutions per site is greater than 0.2, the improved estimator provides a much better estimator for the variance compared with the conventional one.

For the two-parameter model, Table 2 provides the simulation results for different transition/transversion ratios. It can be seen that the improved estimator outperforms the conventional estimator.

Although many more sophisticated methods for estimating the number of nucleotide substitutions per site between two sequences ($K$) are available, the one- and two-parameter methods are still very widely used. In addition, the two-parameter method is used in Li et al. (1985), Li (1993) and Ina (1995) for estimating the number of substitutions per synonymous site and the number of substitutions per nonsynonymous site, and the method by Li

**Table 1**
Comparison of the conventional estimator $V(K)$ and the new estimator $V^*(K)$ for the one-parameter model

| Sequence length ($L$) | Expected number of substitutions per site | True variance | Estimator | |
|---|---|---|---|---|
| | | | $V(K)$ | $V^*(K)$ |
| 500 | 0.1 | 0.000362595 | 0.000219929 | 0.000311769 |
| | 0.2 | 0.001404189 | 0.000494168 | 0.001479758 |
| | 0.3 | 0.004145225 | 0.000830774 | 0.004766365 |
| | 0.4 | 0.010529535 | 0.001247488 | 0.012617401 |
| | 0.5 | 0.025591656 | 0.001776986 | 0.029761526 |
| | 0.6 | 0.061907183 | 0.002434374 | 0.063117721 |
| | 0.7 | 0.141074137 | 0.003261173 | 0.123747284 |
| 1000 | 0.1 | 0.000196551 | 0.000110974 | 0.0001567 |
| | 0.2 | 0.000716157 | 0.000247913 | 0.000735775 |
| | 0.3 | 0.00203738 | 0.000416255 | 0.002360282 |
| | 0.4 | 0.005212835 | 0.000625914 | 0.00626774 |
| | 0.5 | 0.013088052 | 0.000886346 | 0.014583643 |
| | 0.6 | 0.03068909 | 0.001209791 | 0.030645552 |
| | 0.7 | 0.073055629 | 0.001617717 | 0.059929941 |
| 5000 | 0.1 | 3.84488E-05 | 2.21349E-05 | 3.08939E-05 |
| | 0.2 | 0.000145207 | 4.93066E-05 | 0.000143897 |
| | 0.3 | 0.000403367 | 8.27823E-05 | 0.000460616 |
| | 0.4 | 0.000997212 | 0.000124208 | 0.001215598 |
| | 0.5 | 0.002500406 | 0.000175767 | 0.002822049 |
| | 0.6 | 0.005804204 | 0.000240054 | 0.005940681 |
| | 0.7 | 0.013515512 | 0.000320543 | 0.011582702 |

**Table 2**
Comparison of the conventional estimator $V_2$ and the new estimator $V_2{}^*$ for the two-parameter model when the ratio of transition/transversion, $k$, is set to be 1, 2 or 5

| $L$ | $d$ | $k = 1$ | | | $k = 2$ | | | $k = 5$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | True variance $\times 10^{-3}$ | Estimator | | True variance $\times 10^{-3}$ | Estimator | | True variance $\times 10^{-3}$ | Estimator | |
| | | | $V_2 \times 10^{-3}$ | $V_2{}^* \times 10^{-3}$ | | $V_2 \times 10^{-3}$ | $V_2{}^* \times 10^{-3}$ | | $V_2 \times 10^{-3}$ | $V_2{}^* \times 10^{-3}$ |
| 500 | 0.1 | 0.4 | 0.2 | 0.3 | 0.4 | 0.2 | 0.3 | 0.4 | 0.2 | 0.3 |
| | 0.2 | 1.4 | 0.5 | 1.7 | 1.6 | 0.5 | 1.7 | 2.0 | 0.5 | 1.7 |
| | 0.3 | 4.2 | 0.8 | 5.7 | 4.7 | 0.9 | 5.8 | 7.4 | 0.9 | 6.1 |
| | 0.4 | 10 | 1.3 | 16 | 13 | 1.3 | 17 | 23 | 1.5 | 18 |
| | 0.5 | 26 | 1.8 | 41 | 35 | 1.9 | 42 | 72 | 2.3 | 45 |
| | 0.6 | 64 | 2.5 | 90 | 91 | 2.6 | 95 | 224 | 3.3 | 104 |
| | 0.7 | 149 | 3.3 | 186 | 237 | 3.6 | 194 | 678 | 4.8 | 220 |
| 1000 | 0.1 | 0.2 | 0.1 | 0.2 | 0.2 | 0.1 | 0.2 | 0.2 | 0.1 | 0.2 |
| | 0.2 | 0.7 | 0.2 | 0.8 | 0.8 | 0.3 | 0.8 | 1.0 | 0.3 | 0.9 |
| | 0.3 | 2.1 | 0.4 | 2.8 | 2.4 | 0.4 | 2.9 | 3.6 | 0.5 | 3.0 |
| | 0.4 | 5.3 | 0.6 | 8.0 | 6.6 | 0.6 | 8.2 | 11 | 0.7 | 8.7 |
| | 0.5 | 13 | 0.9 | 20 | 18 | 0.9 | 20 | 36 | 1.1 | 22 |
| | 0.6 | 31 | 1.2 | 44 | 42 | 1.3 | 45 | 106 | 1.6 | 49 |
| | 0.7 | 75 | 1.6 | 89 | 104 | 1.8 | 92 | 300 | 2.3 | 102 |
| 5000 | 0.1 | 0.03 | 0.02 | 0.03 | 0.04 | 0.02 | 0.03 | 0.04 | 0.02 | 0.03 |
| | 0.2 | 0.1 | 0.05 | 0.2 | 0.2 | 0.05 | 0.2 | 0.2 | 0.05 | 0.2 |
| | 0.3 | 0.4 | 0.08 | 0.6 | 0.5 | 0.08 | 0.6 | 0.7 | 0.09 | 0.6 |
| | 0.4 | 1.0 | 0.1 | 1.5 | 1.2 | 0.1 | 1.6 | 2.1 | 0.1 | 1.6 |
| | 0.5 | 2.5 | 0.2 | 3.8 | 3.1 | 0.2 | 3.9 | 6.1 | 0.2 | 4.1 |
| | 0.6 | 5.8 | 0.2 | 8.4 | 7.9 | 0.3 | 8.6 | 17 | 0.3 | 9.2 |
| | 0.7 | 14 | 0.3 | 17 | 19 | 0.3 | 18 | 51 | 0.5 | 19 |

$d$ denotes the expected number of substitutions per site.

(1993) is commonly used in current literature. Therefore, accurate estimation of the variance of $K$ for the one- and two-parameter methods is desirable. An alternative method used to improve the variance estimator in the literature is the bootstrap approach. However, since this approach does not provide a closed form for the variance, it requires heavier computations than do the improved variance estimators we derived in this paper. Our estimators have closed forms, so they can be easily applied or included in a computational package such as MEGA4.

In conclusion, the proposed new variance estimators provide substantial improvements for the variance estimation. A computer program for the present variance estimators is available from the author upon request. Online calculations are available at the website: http://cg1.iis.sinica.edu.tw/~var-esti/.

## Acknowledgments

This study was supported by Academia Sinica, Taiwan and NIH Grant GM30998.

## References

Holmquist, R., 1971. Theoretical foundations for a quantitative approach to paleogenetics. Part I: DNA. J. Mol. Evol. 1, 115–133.

Ina, Y., 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 40, 190–226.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Kaplan, N., Risko, K., 1982. A method for estimating rates of nucleotide substitution using DNA sequence data. Theor. Popul. Biol. 21, 318–328.

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.

Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78, 454–458.

Kimura, M., Ohta, T., 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. 2, 87–90.

Lanave, C., Preparata, G., Saccone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20, 86–93.

Li, W.H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

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.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.