Journal of Theoretical Biology 254 (2008) 164–167

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 α, among the four nucleotide types. Since the time of divergence between two sequences is usually unknown, we cannot estimate α 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

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

The sampling variance is approximately given by

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

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

$$-\frac{3}{4}1n\left(1-\frac{4}{3}\frac{X}{L}\right) \approx -\frac{3}{4}1n\left(1-\frac{4}{3}p\right) + \left(\frac{X}{L}-p\right)/(1-(4/3)p) + \left(\frac{X}{L}-p\right)^2 2/(3(1-(4/3)p)^2) \quad (5)$$

From the formula

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

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

$$Var(K) = E\left[\left(-\frac{3}{4}1n\left(1-\frac{4}{3}\frac{X}{L}\right)\right)^2\right] - \left[E\left(-\frac{3}{4}1n\left(1-\frac{4}{3}\frac{X}{L}\right)\right)\right]^2 \quad (6)$$

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

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

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

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

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

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

we have

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

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:

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

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

$$f(X_1,X_2) \approx -\frac{1}{2}\ln(1-2P-Q) - \frac{1}{4}\ln(1-2Q) + (X_1-PL)\frac{1}{L(1-2P-Q)} + (X_2-QL)\frac{1}{2L}\left(\frac{1}{1-2P-Q}+\frac{1}{1-2Q}\right) + \frac{1}{2}\left\{(X_1-PL)^2\frac{1}{L^2}\frac{2}{(1-2P-Q)^2} + 2(X_1-PL)(X_2-QL)\frac{1}{L^2(1-2P-Q)^2} + (X_2-QL)^2\frac{1}{L^2}\left(\frac{1}{2(1-2P-Q)^2}+\frac{1}{(1-2Q)^2}\right)\right\} \quad (11)$$

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

$$V(K_2) \approx \frac{1}{L}\left[P\left(\frac{1}{1-2P-Q}\right)^2 + Q\left(\frac{1}{2-4P-2Q}+\frac{1}{2-4Q}\right)^2 - \left(\frac{P}{1-2P-Q}+\frac{Q}{2-4P-2Q}+\frac{Q}{2-4Q}\right)^2\right] + S \quad (12)$$