Molecular Phylogenetics and Evolution 60 (2011) 472–479

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Confidence intervals for the substitution number in the nucleotide substitution models

Hsiuying Wang

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 5 February 2011
Revised 5 May 2011
Accepted 18 May 2011
Available online 26 May 2011

Keywords:
One-parameter model
Two-parameter model
Confidence interval
Coverage probability
Binomial distribution
Substitution rate

ABSTRACT

In the nucleotide substitution model for molecular evolution, a major task in the exploration of an evolutionary process is to estimate the substitution number per site of a protein or DNA sequence. The usual estimators are based on the observation of the difference proportion of the two nucleotide sequences. However, a more objective approach is to report a confidence interval with precision rather than only providing point estimators. The conventional confidence intervals used in the literature for the substitution number are constructed by the normal approximation. The performance and construction of confidence intervals for evolutionary models have not been much investigated in the literature. In this article, the performance of these conventional confidence intervals for one-parameter and two-parameter models are explored. Results show that the coverage probabilities of these intervals are unsatisfactory when the true substitution number is small. Since the substitution number may be small in many situations for an evolutionary process, the conventional confidence interval cannot provide accurate information for these cases. Improved confidence intervals for the one-parameter model with desirable coverage probability are proposed in this article. A numerical calculation shows the substantial improvement of the new confidence intervals over the conventional confidence intervals.

1. Introduction

A basic process in the evolution of DNA sequences is the substitution of one nucleotide for another during evolution. But since the substitution of one allele for another in a population generally takes thousands of years or longer to complete, the process cannot be directly observed. Thus, 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 the Jukes and Cantor (1969) one-parameter model and the Kimura (1980) two-parameter model (Graur and Li, 1999). For a DNA sequence, the Jukes and Cantor one-parameter model assumes that substitutions occur with equal probability, say α, among the four nucleotide types, A, T, C, G. 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:

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

where p is the probability that the two sequences are different at a site at time t. They proposed the estimator

$$K_1 = -\frac{3}{4}\ln\left(1-\frac{4}{3}\hat{p}\right) \tag{2}$$

to estimate K, where $\hat{p} = X/L$ is the observed proportion of different nucleotides between the two sequences.

The variance of K can be approximated by

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

E-mail address: wang@stat.nctu.edu.tw

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.05.013

By Kimura and Ohta (1972), an estimator for the variance of K is

$$V(K_1) = \frac{\hat{p} - \hat{p}^2}{L(1 - \frac{4}{3}\hat{p})^2}.$$

Although the Jukes and Cantor model is a simple model and many substitution models have been constructed in the literature to compete with it, it is still widely-used due to its simplicity and adaptability for many applications (Fu, 1995; Wirgart et al., 1998; Rosenberg, 2005; Chor et al., 2006, etc.).

In the case of the two-parameter model, the differences between two sequences are classified into transitions and transversions. Transitions are changes between A and G (purines) or between C and T (pyrimidines). Transversions are changes between a purine and a pyrimidine. The substitute probabilities of transition and transversion are assumed to be different. 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 sequences. By Kimura (1980), the number of nucleotide substitutions per site between the two sequences, K_2, is estimated by

$$K_2 = \frac{1}{2}\ln\left(\frac{1}{1 - 2\widehat{P} - \widehat{Q}}\right) + \frac{1}{4}\ln\left(\frac{1}{1 - 2\widehat{Q}}\right)$$

and the sampling variance is approximately given by

$$V(K_2) = \frac{1}{L}\left(\widehat{P}\left(\frac{1}{1 - 2\widehat{P} - \widehat{Q}}\right)^2 + \widehat{Q}\left(\frac{1}{2 - 4\widehat{P} - 2\widehat{Q}} + \frac{1}{2 - 4\widehat{Q}}\right)^2 - \left(\frac{\widehat{P}}{1 - 2\widehat{P} - \widehat{Q}} + \frac{\widehat{Q}}{2 - 4\widehat{P} - 2\widehat{Q}} + \frac{\widehat{Q}}{2 - 4\widehat{Q}}\right)^2\right). \quad (3)$$

Since the above two variance estimators underestimate the true variances in most mircumstances, Wang et al. (2008) derive improved variance estimators using a higher-order Taylor expansion and empirical methods.

The above illustration is under the assumption that the rate of nucleotide substitution is the same for all nucleotide sites. However, this assumption may not hold in some situations because the nucleotide sequences have functional constraints and usually form a secondary structure consisting of loops and stems that have different substitution rates. Kocher and Wilson (1991), Tamura and Nei (1993) and Wakeley (1993, 1994) suggest that the substitution rate varies from site to site according to the gamma distribution for this case.

When the nucleotide substitution at each site follows the Jukes and Cantor model but the substitution rate 3α varies with the gamma distribution $\Gamma(a, b)$, by Golding (1983) and Nei and Gojobori (1986), the expected number of substitutions per site becomes

$$H = \frac{3}{4}a\left[\left(1 - \frac{4}{3}p\right)^{-1/a} - 1\right]$$

and the variance for the number of the substitutions per site is

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

where a is the shape parameter of the gamma distribution with the density function $f(x) = [b^a/\Gamma(a)]e^{-bx}x^{a-1}$. Note that H and $V(H)$ depend on only one parameter of the gamma distribution, but not the two parameters a and b because a/b is the mean of the substitution rate 3α, and b is a function of a and α (Nei and Gojobori, 1986).

The estimators

$$H_1 = \frac{3}{4}a\left[\left(1 - \frac{4}{3}\hat{p}\right)^{-1/a} - 1\right]$$

and

$$V(H_1) = \frac{\hat{p}(1-\hat{p})}{L}\left[\left(1 - \frac{4}{3}\hat{p}\right)^{-2(1/a+1)}\right]$$

are used to estimate H and $V(H)$.

The true number of substitutions per site is usually approximated by point estimators. However, from a statistical point of view, a better approach is to report a confidence interval of the substitution number instead of a point estimator for the number of substitutions because the point estimation can only provide a rough estimate without any information about its precision. The confidence interval estimation can quantify the uncertainty associated with the estimate such that we can have the confidence degree if the true K or H is belonged to the intervals. In many studies, the confidence intervals are reported associated with the point estimation, e.g. Yang (2007). The conventional confidence interval is constructed using the normal approximation, which can achieve the desirable coverage probability when the sample size is large enough. Here the sample size is the length of a sequence. However, even when the length is large, from the study shown in Section 2, the conventional confidence interval suffers from the serious drawback of unsatisfactory coverage probability when the true substitution number is small. Since in the evolutionary process of DNA sequences, the true substitution number per site may be very small, the behavior of a confidence interval for the small substitution number case is especially important. Accordingly, the information provided by the conventional confidence interval is not very accurate.

In this article, modified confidence intervals for the one-parameter model with more satisfactory coverage probability are proposed in both constant substitution rate and variable substitution rate models. These modified confidence intervals are constructed from a modification approach used in the literature for constructing confidence intervals of a binomial proportion.

This article is organized as follows. The coverage probability and expected length of the conventional confidence interval for substitution models with constant substitution rate and variable substitution rate are shown in Section 2. Section 3 gives the proposed confidence intervals as well as their performances. Section 4 provides an algorithm for selecting a factor such that the coverage probability of the confidence interval is close to a desirable level. The proposed methods are illustrated by a real data example analyzing the substitution number of owlet-nightjars species in Section 5. The article concludes in Section 6 with a summary.

2. The existing methods

We first consider the constant substitution rate case. A statistical interval $(L(\hat{p}), U(\hat{p}))$ is said to be a level $1-\gamma$ confidence interval of K if it can cover the true K with at least probability $1-\gamma$, which is defined as

$$P_K(L(\hat{p}) < K < U(\hat{p})) \geqslant 1 - \gamma.$$

The probability $P_K(L(\hat{p}) < K < U(\hat{p}))$ is the coverage probability of the confidence interval, and $E_K(|U(\hat{p}) - L(\hat{p})|)$ is the expected length of the confidence interval under the true substitution number K. Usually the coverage probability of a level $1-\gamma$ confidence interval we constructed based on the normal approximation may not be equal to $1-\gamma$ and may be close to the nominal level $1-\gamma$ only when the sample size is large. Furthermore, its coverage probability