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

could be far away from the nominal level $1-\gamma$ for a fixed sample size. We usually evaluate confidence intervals in terms of their coverage probabilities. A confidence interval with coverage probability close to the nominal level is regarded as better than a confidence with coverage probability far away from the nominal level.

In the Jukes and Cantor one-parameter model, when the substitution rate is assumed to be the same for all sites, the conventional confidence interval is constructed by the normal approximation that the statistic

$$\frac{K_1 - K}{\sqrt{V(K_1)}} \tag{4}$$

has an asymptotic standard normal distribution. By inverting from the probability

$$P\left(\frac{|K_1 - K|}{\sqrt{V(K_1)}} < z_{1-\frac{\gamma}{2}}\right) = 1-\gamma, \tag{5}$$

we have the level $1-\gamma$ conventional confidence interval

$$(K_1 - z_{1-\frac{\gamma}{2}}\sqrt{V(K_1)}, K_1 + z_{1-\frac{\gamma}{2}}\sqrt{V(K_1)}), \tag{6}$$

where $z_{1-\gamma/2}$ is the upper $1-\gamma/2$ cutoff point of the standard normal distribution.

Note that the equality in (5) only holds when L is large enough. Consequently, the coverage probability of (6) is not exactly equal to $1-\gamma$ for a fixed sample size. We explore its coverage probability by a simulation study.

When the substitution rate is assumed to follow the gamma distribution with shape parameter a, a confidence interval constructed by the usual normal approximation uses the fact that the statistic

$$\frac{H_1 - H}{\sqrt{V(H_1)}} \tag{7}$$

has an asymptotic standard normal distribution. Consequently, we have the level $1-\gamma$ conventional confidence interval

$$(H_1 - z_{1-\frac{\gamma}{2}}\sqrt{V(H_1)}, H_1 + z_{1-\frac{\gamma}{2}}\sqrt{V(H_1)}) \tag{8}$$

We conduct a simulation to explore their coverage probability. For the constant substitution rate case, the simulation method is to generate two descendant sequences from an ancestral sequence. First we set a value for αt, say v_0, then generate descendant sequences with probability $1/4 + 3/4e^{-4\alpha t}$ that the nucleotide at a site in a descendant sequence is the same as that in an ancestral sequence and with probability $1/4 - 1/4e^{-4\alpha t}$ that the nucleotide at a site in a descendant sequence is equal to one of the three other bases from the ancestral sequence. Then we compute the proportion of the different nucleotide in these two descendant sequences and use the proportion as $\hat{p}$ to derive K_1 and $V(K_1)$. To derive the coverage probability of the confidence interval, it is necessary to calculate the proportion if the true $K = 2(3\alpha t)$ belongs to the interval from a simulation study. We replicate the above process 1000 times in generating the sequences and deriving K_1, and then calculate the proportion that the interval based on K_1 covers the true K.

The proportion is the coverage probability of the confidence interval approximated by the simulation. The simulation for the Kimura two-parameter model is to generate the descendant sequences from a common origin using the probability setup for the two-parameter model. By using a method similar to the one-parameter model, we can derive the coverage probability for the two-parameter model. The model and simulation approach are referred to Graur and Li (1999), Nei and Kumar (2000) and Yang (2007).

For the variable substitution rate case, the simulation method is that for each site, we generate $\alpha' = 3\alpha$ value from a gamma distribution $\Gamma(a, b)$ for each site, then use these $\alpha'/3$ values as the substitution rate to generate two descendant sequences from an ancestral sequence. Then by a similar argument as the constant substitution rate case, the coverage probability of a confidence interval for $H = 2(a/b)$ can be calculated.

Figs. 1 and 2 plot the coverage probability and expected length of the confidence interval (6) corresponding to different K values, which shows that the coverage probability of the conventional confidence interval is much lower than the nominal level when $L = 100$ and 500 for $K \leqslant 0.06$.

In the Kimura two-parameter model, by an argument similar to that in the one-parameter model, the $1-\gamma$ level confidence interval of K_2 is

$$(K_2 - z_{1-\gamma/2}\sqrt{V(K_2)}, K_2 + z_{1-\gamma/2}\sqrt{V(K_2)}). \tag{9}$$

Fig. 3 shows the coverage probabilities of the confidence intervals for $L = 100$ and 500 when the expected substitution number per site is less than 0.11.

The coverage probability, increasing as K increases, of the interval (6) for the Jukes Cantor model is much lower than the nominal level 0.95 when the true K is small. Note that the coverage probability is not a very smooth curve of K shown in Fig. 1 because X is a discrete random variable which leads to the oscillation of the coverage probability. We can see that the performance of the conventional confidence interval is not satisfactory because its coverage probability cannot reach the nominal level. This is the same as the Kimura two-parameter model, where the coverage probability is much lower than the nominal level 0.95.

In fact, the performance of the conventional confidence interval for the two-parameter model is even worse. For the one-parameter model, Fig. 1 shows that the coverage probability increases to 0.95 when L increases. However, from Fig. 3, the deviation of the coverage probability to the nominal level 0.95 multiplies when L increases. This indicates that the confidence interval constructed for the two-parameter model is less satisfactory in estimating the

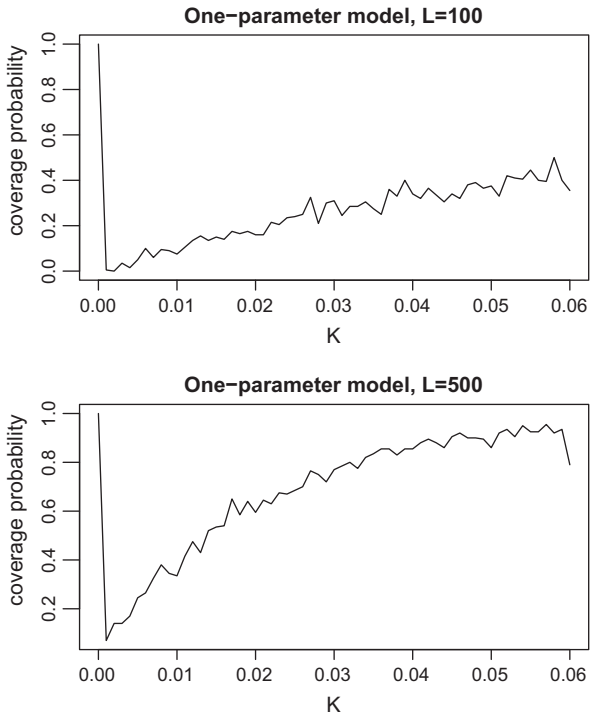


Fig. 1. The coverage probabilities of the level 0.95 confidence intervals (6) when $L = 100$ and 500 for $K \leqslant 0.06$.

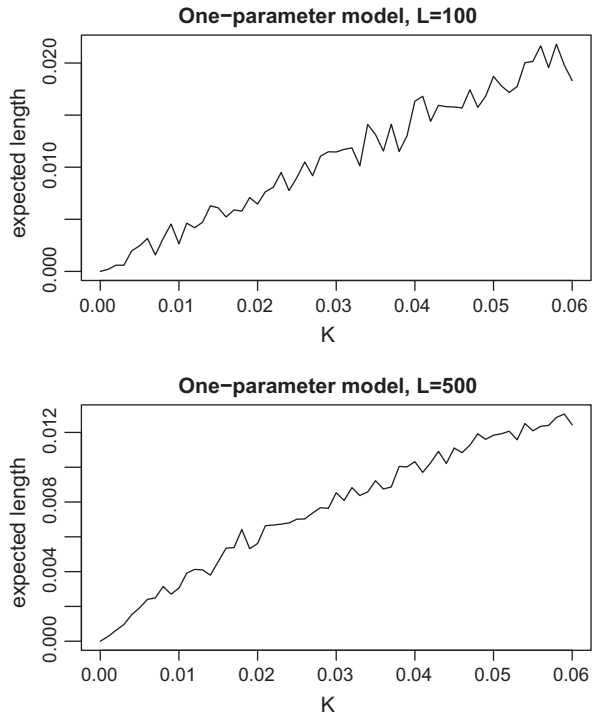


Fig. 2. The expected lengths of the level 0.95 confidence intervals (6) when $L = 100$ and 500 for $K \leqslant 0.06$.

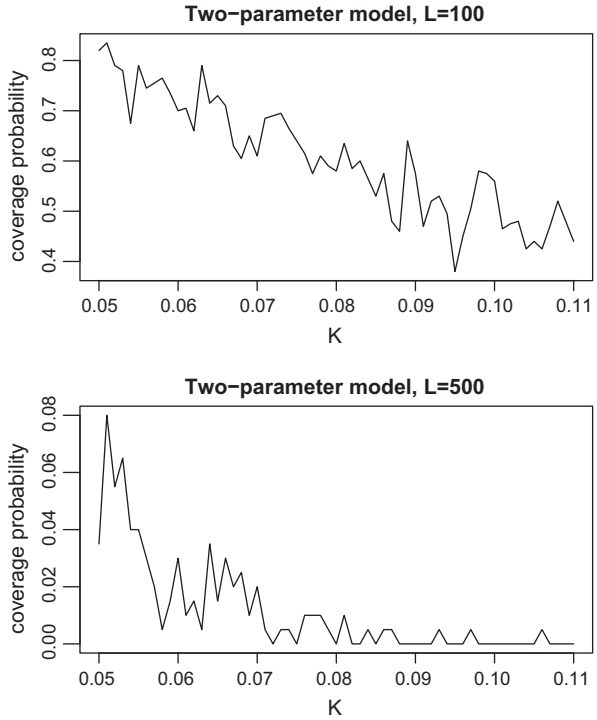


Fig. 3. The coverage probabilities of the level 0.95 confidence intervals (9) for $L = 100$ and 500 for $K \leqslant 0.11$. Here the substitution probability of transition is fixed at 0.05, and of the substitution probability of transversion ranges from 0 to 0.03.

true K. This drawback may be due to the bias of the estimator K_2. Accordingly, we mainly focus on constructing improved confidence intervals based on the one-parameter model in this article.

It is worth noting that there are other approaches such as the bootstrap method and jackknife method for constructing confidence intervals in phylogenetic study (Dopazo, 1994, etc.). However, these methods cannot provide a closed form. In this study, we aim to establish improved closed form intervals such that they can be easily implemented in real applications.

3. New confidence interval

It is not surprising that the coverage probability of the conventional confidence interval goes to zero as K goes to zero. This phenomenon also commonly occurs in a simple model, the binomial distribution (see Agresti and Coull, 1998; Wang, 2007). In the one-parameter and two-parameter models, the parameters of interest are functions of the binomial proportion p. Since there are several approaches in the literature to modify the conventional confidence interval such that the coverage probability for small p can be improved, we extend these approaches to the one-parameter model.

Both models for the constant substitution rate and variable substitution rate are considered.

3.1. Constant substitution rate

First we deal with the model that the substitution rate is assumed to be the same for each site.

The first method is the extension from the score approach. The score $1-\gamma$ confidence interval for the binomial proportion of a binomial distribution is constructed by inverting from the set $\{p : |x/n - p|/\sqrt{p(1-p)/n} \leqslant z_{1-\gamma/2}\}$. The endpoints of the score confidence interval are the two solutions of p found by solving the equation $(x/n-p)^2/(p(1-p)/n) = z^2_{1-\gamma/2}$. By a similar argument, the score interval for K is obtained by inverting from the set

$$\{|K_1 - K|/\sqrt{V(K)} \leqslant z_{1-\gamma/2}\}. \tag{10}$$

Note that by (1), we have

$$p = 3/4(1 - e^{-4/3K}). \tag{11}$$

Replacing it in $V(K)$, we have

$$V(K) = \frac{3(-3 + 2e^{4K/3} + e^{8K/3})}{16L}. \tag{12}$$

The score confidence interval can be derived by replacing (12) in (10) and then solving the equation

$$(K_1 - K)^2/V(K) = z^2_{1-\gamma/2} \tag{13}$$

in K. However, the above equation does not have a closed form, which would need to be solved by a numerical calculation. To prevent this disadvantage, which may cause the inconvenience of usage of the interval, we can use the Taylor expansion to approximate the terms $e^{4K/3}$ and $e^{8K/3}$ in (12) by $1 + 4K/3$ and $1 + 8K/3$ because the substitute number should be small. Consequently, the variance can be approximated by K/L. Replacing the approximated variance in (13) and solving the equation, we have the approximated $1-\gamma$ score confidence interval

$$\left(\frac{2K_1L + z^2_{1-\gamma/2} - z_{1-\gamma/2}\sqrt{4K_1L + z^2_{1-\gamma/2}}}{2L}, \frac{2K_1L + z^2_{1-\gamma/2} + z_{1-\gamma/2}\sqrt{4K_1L + z^2_{1-\gamma/2}}}{2L}\right). \tag{14}$$

The second approach is an extension of the Agresti and Coull approach (1998) to modify the confidence interval (6) by replacing X and L by $X + z^2_{1-\gamma/2}/2$ and $L + z^2_{1-\gamma/2}/2$, respectively. Let

$$\hat{\hat{p}} = \frac{X + z^2_{1-\gamma/2}/2}{L + z^2_{1-\gamma/2}}, K_1^c = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{\hat{p}}\right) \quad \text{and } V^c(K_1)$$
$$= \frac{\hat{\hat{p}} - \hat{\hat{p}}^2}{L\left(1 - \frac{4}{3}\hat{\hat{p}}\right)^2}.$$

The interval is proposed to have the form

$$(K_1^c - z_{1-\gamma/2}\sqrt{V^c(K_1)}, K_1^c + z_{1-\gamma/2}\sqrt{V^c(K_1)}). \quad (15)$$

This approach can successfully increase the coverage probability when K goes to zero. We call this interval an adjusted confidence interval.

To compare the intervals, we conduct a simulation to explore their coverage probabilities and expected lengths, as shown in Figs. 4 and 5.

The comparison of the two proposed level 0.95 confidence intervals of K for $L = 100$ and 500 is shown in Figs. 4 and 5 for small K. The solid and dashed lines represent the coverage probabilities of level 0.95 score and adjusted confidence intervals in Fig. 4. The coverage probability of a good confidence interval should be very close to 0.95 for any K. The simulation results show that the new confidence intervals substantially improve the coverage probability when K is small, compared with the conventional confidence interval. We also conduct a simulation for larger, L such as $L = 2000$. They have similar performance as the above cases that the coverage probability of the adjusted interval is higher than the score interval.

The comparison of the expected lengths of the three confidence intervals shows the conventional confidence interval has the shortest expected length. In the criterion of evaluating a confidence interval, although we prefer a confidence interval with shorter expected length, the most important thing is to evaluate its coverage probability performance. A conventional interval, with a smaller expected length, has poor coverage probability because it cannot cover the true K with desirable frequency.

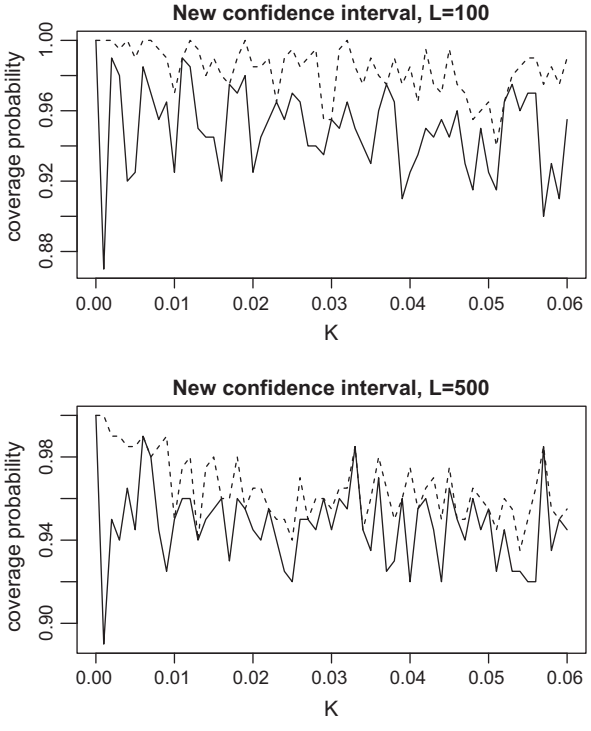


Fig. 4. Solid and dashed lines present the coverage probabilities for the level 0.95 score and adjusted confidence intervals, respectively when $L = 100$ and $L = 500$ for $K \leqslant 0.06$ for the constant substitution rate case.

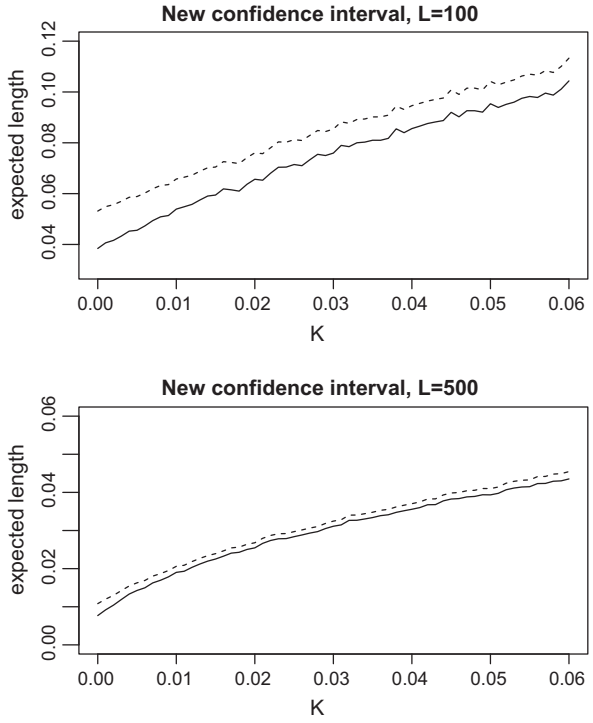
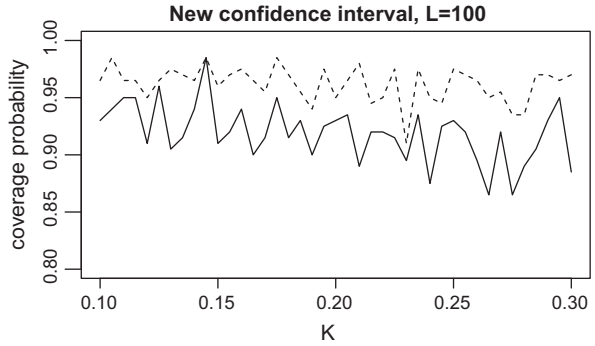


Fig. 5. Solid and dashed lines present the expected lengths for the level 0.95 score and adjusted confidence intervals, respectively when $L = 100$ and $L = 500$ for $K \leqslant 0.06$ for the constant substitution rate case.

For the performance of the two new intervals, the simulation study shows that the adjusted interval always has higher coverage probability and longer expected length than the score confidence interval. In this case, we would prefer the score confidence interval because its average coverage probability is closer to the nominal level when the true K is small.

When the true K is not small, the performance of the two new intervals is shown in Fig. 6. The coverage probability of the adjusted interval is close to the nominal level 0.95, but the coverage probability of the score interval is lower than 0.95. The performance of the score interval is worse than the adjusted interval when the true K is not small. We recall the method used to construct (14) is to approximate $e^{4K/3}$ and $e^{8K/3}$ by $1 + 4K/3$ and $1 + 8K/3$, respectively, in (12) and (13). The approximation is only appropriate when K is small. When K is not small, to derive a more accurate score interval, a better approximation for $e^{4K/3}$ is using a Taylor expansion up to the second order term $1 + 4K/3 + (4K/3)^2/2$ or up to a higher order term. We also can derive the score interval by a numerical method to solve (13). Thus, when we do not have information about the range of the true K, we recommend the adjusted interval, or using a more accurate score interval.

Note that according to Wang et al. (2008), there exist better variance estimators for the variance of K in the one-parameter and two-parameter models instead of the variance estimators $V(K_1)$ and $V(K_2)$. Consequently, we can replace the usual variances in (6), (9) by the improved variances proposed in Wang et al. (2008) to construct another confidence interval. However, this method cannot substantially improve the coverage probability when the true K is small from a simulation study. Thus, to improve the coverage probability, we propose the score approach and the adjusted approach to construct new confidence intervals.



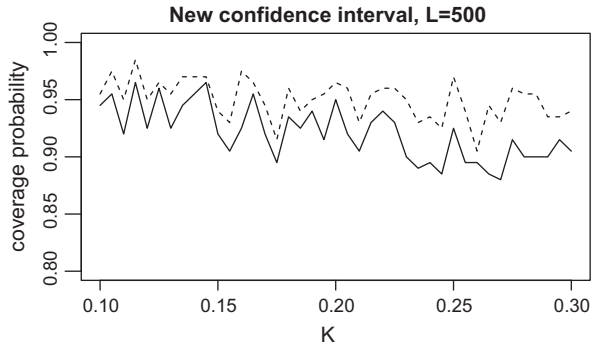
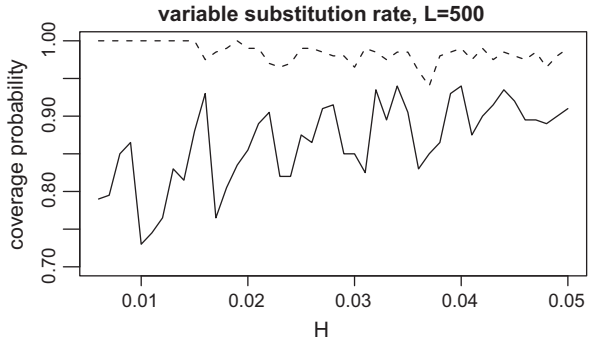


Fig. 6. Solid and dashed lines present the expected lengths for the level 0.95 score and adjusted confidence intervals, respectively when $L = 100$ and $L = 500$ for $K \leqslant 0.3$ for the constant substitution rate case.



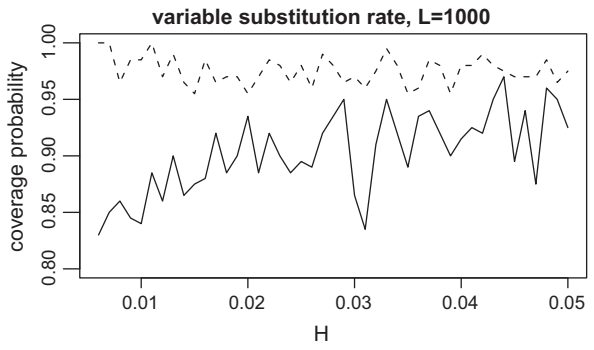


Fig. 7. Solid and dashed lines present the coverage probabilities for the level 0.95 standard and adjusted confidence intervals, respectively when $L = 500$ and $L = 1000$ for $H \leqslant 0.05$ for the variable substitution case.

3.2. Variable substitution rate

We proceed to consider the Juke and Cantor model when the substitution rate is assumed to follow a gamma distribution $\Gamma(a, b)$ with a shape parameter a.

For this case, applying the score approach leads to a messy formula for the confidence interval of H because the endpoints of the interval are approximated by solving a polynomial equation with degree 3 after simplification. Although it is hard to provide a closed form for the score interval of H, it is feasible to derive the confidence interval by numerical calculation.

In this case, the second approach, the adjusted approach, may be simpler and more useful here. The adjusted confidence interval is constructed by replacing H_1 and $V(H_1)$ in (8) as

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

and

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

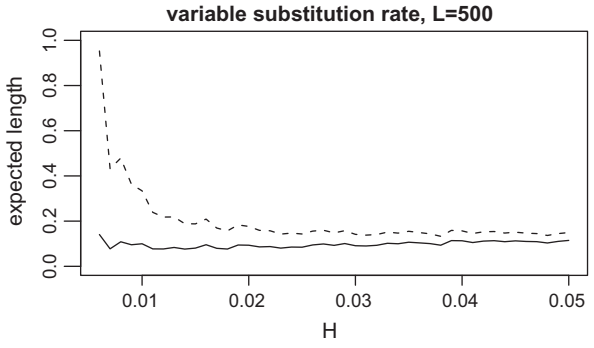
respectively.

This leads to the confidence interval

$$(H_1^c - z_{1-\frac{\gamma}{2}}\sqrt{V(H_1^c)}, H_1^c + z_{1-\frac{\gamma}{2}}\sqrt{V(H_1^c)}). \quad (16)$$

To compare the performance of the intervals, we conduct a simulation study to compare their coverage probabilities and expected lengths, which are shown in Figs. 7 and 8.

The comparisons of the standard and proposed level 0.95 confidence intervals of H for $L = 500$ and 1000 are shown in Figs. 7 and 8. Here we consider the longer sequence lengths $L = 500$ and 1000 instead of $L = 100$ and 500 in the constant substitution rate case because the estimation for the variable substitution rate case is more unstable than that for the constant substitution rate case. The solid and dashed lines represent the coverage probabilities of level 0.95 standard and adjusted confidence intervals in Fig. 7 for



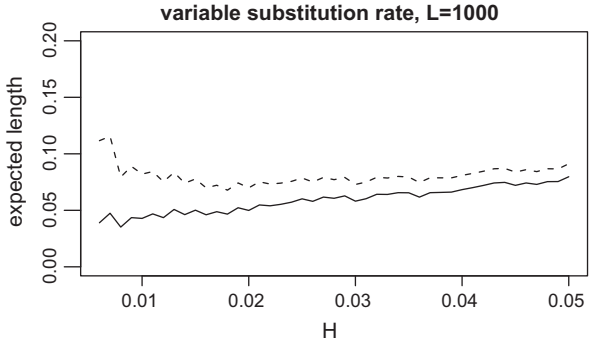


Fig. 8. Solid and dashed lines present the expected lengths for the level 0.95 standard and adjusted confidence intervals, respectively when $L = 500$ and $L = 1000$ for $H \leqslant 0.05$ for the variable substitution case.

Table 1
The average coverage probabilities of the level 95% conventional interval and adjusted interval for $0 < K < 0.06$ in the variable substitution model.

L	Conventional interval	Adjusted interval
500	0.858	0.985
1000	0.891	0.975
2000	0.913	0.97

Table 2
The level 95% conventional, score and adjusted confidence intervals for sequences AY090696 and AY09069.

Conventional confidence interval	(0.0001, 0.0145)
Score confidence interval	(0.0028, 0.01875)
Adjusted confidence interval	(0.0021, 0.0195)

$H \leqslant 0.05$. The coverage probability of the standard confidence interval is clearly lower than the nominal level, which reveals the same disadvantage as in the constant substitution rate case. The adjusted confidence interval has higher coverage probability. Although it seems higher than the nominal level, we still prefer the adjusted confidence interval because its minimum coverage probability is near the nominal level 0.95. As we expect, the expected length of the adjusted confidence interval is longer than that of the standard confidence interval because the adjusted confidence interval has higher coverage probability. The simulation results of $L = 500$ and $L = 1000$ reveal that the intervals has better performance when the length L increases.

The average coverage probabilities of the proposed interval and the conventional interval are included in Table 1 to support the advantages of the proposed confidence intervals. Table 1 shows the coverage probabilities of the level 0.95 conventional interval and the level 0.95 adjusted interval for K in (0, 0.06) in the variable substitution model for different lengths. It reveals that the average coverage probability of the adjusted interval is closer to the nominal level 0.95 than the conventional interval.

4. Selection of γ value

Although the coverage probabilities for the new confidence intervals can be substantially improved over the conventional interval when the true K or H is small, Figs. 4–7 show that they still cannot reach the nominal level 0.95 for all p. Since the conventional and new confidence intervals are constructed based on the normal approximation, when the sample size is not large enough, their performance may not be good enough. To prevent this problem from occurring at the small sample size case, we propose an algorithm that is modified from algorithms in Wang (2007, 2009) to select an appropriate γ values in (14)–(16) such that the minimum coverage probability or the average coverage probability can reach a desirable level. The minimum coverage probability denotes the minimum value of the coverage probability of a confidence interval when K ranges over the domain. The average coverage probability denotes the average value of the coverage probability with respect to a prior for K under the Bayesian setup. The details are referred to Wang (2009).

For the model with constant substitution rate, we can first consider the minimum coverage probability criterion. To construct a modified confidence interval by applying the algorithms in Wang (2007, 2009), first we transform the confidence interval of K to the confidence interval of p. By (1), K is a one-to-one function of p, where p can be expressed as (11) in terms of K. Thus, for a level $1-\gamma$ confidence interval $(L(X), U(X))$ of K, we can make a transformation such that the interval $(3/4(1-e^{-4/3L(X)}), 3/4(1-e^{-4/3U(X)}))$ is a level $1-\gamma$ confidence interval of p. If we hope to obtain a confidence interval with a minimum coverage probability 0.95, then we can select an appropriate $z_{1-\gamma/2}$ by using a procedure modified from Wang (2007).

The algorithm of the modified procedure in Wang (2007) is as follows. Since the length of the sequence is L, the possible values of observed X are $\{0, 1, \ldots, L\}$. To adopt the procedure, the confidence interval needs to satisfy the assumption that $3/4(1-e^{-4/3L(X)})$ and $3/4(1-e^{-4/3U(X)})$ are increasing functions of X (Wang, 2010). In addition to this increasing condition, from (2), since the log function must define on a positive domain, it needs to require that X/L is less than 3/4. Thus, we have to check if $3/4(1-e^{-4/3L(X)})$ and $3/4(1-e^{-4/3U(X)})$ are increasing functions of X for $X \leqslant 3/4L$. By checking the three confidence intervals (6, 14, and 15) in the one-parameter model, we found that only the score interval (14) satisfies the assumption. Therefore, we modified the score interval using the following procedure to select a γ value in the confidence interval (14) such that the confidence interval has minimum coverage probability 0.95.

Algorithm 1. Procedure to derive a factor in a score confidence interval $(L(X), U(X))$ with the form (14) such that the confidence interval has a minimum coverage probability 0.95.

Step 1. Let $(L^*(X), U^*(X)) = (3/4(1-e^{-4/3L(X)}), 3/4(1-e^{-4/3U(X)}))$.
Step 2. Calculate the endpoints $L^*(X)$ and $U^*(X)$ for $x \in \{0, 1, \ldots, [3/4L]\}$, where $[w]$ denotes the largest integer less than w. Then list the endpoints $L^*(X)$ and $U^*(X)$ that are greater than zero and smaller than 1.
Step 3. Calculate the coverage probabilities for p in the set of endpoints of Step 2 which are greater than zero and smaller than 1. The minimum value of these coverage probabilities is the minimum coverage probability of the confidence interval. The details refer to Wang (2007, 2009).
Step 4. Find the γ value such that the minimum coverage probability derived in Step 3 is equal to 0.95. Then the confidence interval $(L(X), U(X))$ with this γ value is the confidence interval with a minimum coverage probability 0.95.

The confidence level 0.95 in the algorithm can be replaced by other values depending on the precision we prefer.

For a model with variable substitution rate, we can consider the average coverage probability criterion. An argument similar to that in Wang (2009) can be used to construct a confidence interval of H such that it has the desirable average coverage probability with respect to a prior.

5. Illustrative example

We use the DNA sequences of the avian family Aegothelidae (commonly known as owlet-nightjars) to illustrate the proposed approaches. Owlet-nightjars are small nocturnal birds related to the nightjars and frogmouths. Most are native to New Guinea, but some species extend to Australia, Moluccas, and New Caledonia. There is a single monotypic family Aegothelidae with the genus Aegotheles. The family Aegothelidae comprises only nine extant species, all in a single genus, Aegotheles.

Dumbacher et al. (2003) used mitochondrial DNA sequence to construct a phylogeny of the owlet-nightjars. They analyzing mtDNA sequences cytochrome b and ATPase subunit 8 suggests that there are 11 living species of owlet-nightjar and one that went extinct early in the second millennium AD. The taxon listed in Table 2 of Dumbacher et al. (2003) includes albertisi albertisi, wallacii wallacii, wallacii gigas, etc. The Genbank numbers for the sequences are AY090664-AY090698 (for cytochrome b) and AY090699-AY090736 (for ATPase 8).

We use the DNA sequences AY090696 and AY090697 of wallacii wallacii and wallacii gigas to illustrate the new confidence intervals for the one-parameter model. First we apply the multiple sequences alignment procedure of MEGA software to these sequences, and then count the number of different nucleotides between the sequences. The number of different nucleotides between the two sequences is four among 550 which is the length size. The three intervals for the number of nucleotide substitutions per site between the two sequences are listed in Table 2.

Table 2 shows that the lower confidence bounds of the score and adjusted confidence intervals are much larger than those of the conventional confidence interval. This reveals if we adopt the conventional confidence interval, then we may think that it is possible that the number of nucleotide substitutions per site can be near 0.0001. But from the two other intervals, we have 0.95 confidence that the number of nucleotide substitutions per site is greater than 0.002. By the analysis from Sections 2 and 3, when the proportion of the different nucleotide between two sequences is low, the score and adjusted confidence intervals are more reliable than the conventional confidence interval.

6. Conclusion

The performance and construction of confidence intervals for evolutionary models have not been much investigated in the literature. In this article, we explore the conventional confidence interval performance and provide new confidence intervals, which are shown to be better than the conventional confidence interval for estimating the nucleotide substitution number per site when the true number of substitutions is small. Both the constant substitution rate and the variable substitution rate cases are considered. One of the proposed approaches is to extend the score confidence interval for the binomial proportion to construct the improved confidence interval. The other approach is based on an adjusted approach used for the binomial distribution. The simulation results show that the proposed confidence intervals are substantially improved over the standard confidence intervals. For the constant substitution rate case, the two new intervals outperform the conventional interval. For the variable substitution rate case, since the score interval does not have a closed form, the adjusted interval is more feasible for the real applications.

Since the proposed methodologies as well as the conventional approach are based on the normal approximation, when the length of sequences is long, the approaches can perform well. But when the length is not long enough, which can be viewed as a case with small sample size case, the methods are not satisfactory. In this circumstance, a more accurate method is to select an appropriate factor value, which can deal with the shorter length case with high precision.

The confidence interval approach can provide more useful information for estimating the nucleotide substitution number than the point estimation. The approach proposed in this article can provide a more efficient and accurate way in the nucleotide substitution number estimation than the conventional confidence interval.

Acknowledgments

This study was supported by the National Science Council and National Center for Theoretical Sciences in Taiwan.

References

Agresti, A., Coull, B.A., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. Am. Stat. 52, 119–126.
Chor, B., Hendy, M.D., Snir, S., 2006. Maximum likelihood Jukes–Cantor triplets: analytic solutions. Mol. Biol. Evol. 23, 626–632.
Dopazo, J., 1994. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. J. Mol. Evol. 38, 300–304.
Dumbacher, J.P., Pratt, T.K., Fleischer, R.C., 2003. Phylogeny of the owlet-nightjars (Aves: Aegothelidae) based on mitochondrial DNA sequence. Mol. Phylogenet. Evol. 29, 540–549.
Fu, Y., 1995. Linear invariants under Jukes' and Cantor's one-parameter model. J. Theor. Biol. 173, 339–352.
Golding, G.B., 1983. Estimates of sequence divergence: An examination of some assumptions. Mol. Biol. Evol. 1, 125–142.
Graur, D., Li, W.H., 1999 Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, MA.
Holmquist, R., 1971. Theoretical foundations for a quantitative approach to paleogenetics. Part I: DNA. J. Mol. Evol. 1, 115–133.
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.
Kocher, T.D., Wilson, A.C., 1991. Sequence evolution in mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. In: Osawa, S., Honjo, T. (Eds.), Evolution of Life: Fossils, Molecules, and Culture. Springer-Verlag, New York, pp. 391–413.
Lanave, C., Preparata, G., Saccone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20, 86–93.
Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc.
Rosenberg, M.S., 2005. MySSP: non-stationary evolutionary sequence simulation, including indels. Evol. Bioinf. Online 1, 81–83.
Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.
Wakeley, J., 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. 37, 613–623.
Wakeley, J., 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. 11, 436–442.
Wang, H., 2007. Exact confidence coefficients of confidence Intervals for a Binomial Proportion. Stat. Sinica 17, 361–368.
Wang, H., 2009. Exact average coverage probabilities and confidence coefficients of confidence intervals for discrete distributions. Stat. Comput. 19, 139–148.
Wang, H., 2010. Monotone boundary property and the full coverage property of confidence intervals for a binomial proportion. J. Stat. Plan. Infer. 140, 495–501.
Wang, H., Tzeng, Y.H., Li, W.H., 2008. Improved variance estimators for one- and two-parameter models of nucleotide substitution. J. Theor. Biol. 254, 164–167.
Yang, Z., 2007. Computational Molecular Evolution. Oxford University Press.