A Simple Method for Estimating the Parameter of Substitution Rate Variation Among Sites

Xun Gu and Jianzhi Zhang
Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University

When the rate variation among sites is described by a gamma distribution, an important problem is how to estimate the shape parameter $\alpha$, which is an index of the degree of among-site rate variation. The parsimony-based methods for estimating $\alpha$ are simple but biased, i.e., $\alpha$ tends to be overestimated. On the other hand, the likelihood-based methods are asymptotically unbiased but take a huge amount of computational time. In this paper, we have developed a new method to solve this problem: we first estimate the expected number of substitutions at each site, which is corrected for multiple hits, and then estimate the parameter $\alpha$. Our method is computationally as fast as the parsimony method, and the estimation accuracy is much higher than that of parsimony and similar to that of the likelihood method.

Introduction

It is well known that different amino acid residues of a protein may have different functional constraints such that the substitution rate varies among the sites. Although this phenomenon was first described over 20 years ago (Uzzel and Corbin 1971), its importance for molecular evolutionary study has not been recognized until recently. The effect of rate variation among sites on phylogenetic reconstruction and divergence time estimation has also been discussed extensively (e.g., Jin and Nei 1990; Hulsenbeck and Hillis 1993; Tateno, Takezaki, and Nei 1994; Yang 1994, 1996; Miyamoto and Fitch 1996; Gu 1997).

The gamma distribution has been widely used for modeling the rate variation among sites (Uzzel and Corbin 1971; Holmquist et al. 1983; Tamura and Nei 1993; Yang 1993; Gu, Fu, and Li 1995). Under this model, the variation of substitution rate ($\lambda$) among sites can be described as follows:

$$\phi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta \lambda},$$

(1)

where the shape parameter $\alpha$ is important because it describes the degree of rate variation, and $\beta$ is a scalar. Since $1/\sqrt{\alpha}$ is the coefficient of variation of $\lambda$, the larger $\alpha$ is, the weaker the rate variation is, and $\alpha = \infty$ means a uniform rate among sites.

Several methods have been developed for estimating $\alpha$ from sequence data, these methods can be classified into two groups. The first group is the maximum-likelihood (ML) approach, which was constructed under the framework of Felsenstein (1981) (e.g., Yang 1993, Gu, Fu, and Li 1995; Kelly and Rice 1996). However, the algorithms developed by these authors for maximizing the likelihood function are so time-consuming that they are difficult to apply for more than five sequences. Although an approximate method (the discrete-gamma distribution) has been developed (Yang 1994), it is still time-consuming and cannot handle a large amount of sequence data in a short period of time. The second group of methods for estimating $\alpha$ is usually called the parsimony method, which has been widely used because it is computationally fast (e.g., Uzzel and Corbin 1971, Holmquist et al. 1983; Larson 1991; Tamura and Nei 1993; Sullivan, Holsinger, and Simon 1995; Toulouse and Gouy 1997). In these methods, the principle of parsimony (Fitch 1971) was used to infer the (minimum required) number of substitutions. Since the parsimony method tends to underestimate the number of substitutions, it is known that $\alpha$ can be seriously overestimated: in other words, the degree of rate variation among sites can be underestimated (Wakeley 1993). An approach with a combination of likelihood and parsimony methods was proposed by Yang and Kumar (1996). Since their method assumes that all branch lengths are equal in the phylogeny, it may be problematic when branch lengths along the tree are very different.

Large scale sequence analysis calls for methods that can deal with many sequences or many genes at the same time. To complement recent improvement in our computational capacity for DNA sequence analysis, it is also important to study alternative models for developing a statistically unbiased and computationally fast method. In this paper, a simple ML method is proposed which has two steps: (1) at each site, the expected number of substitutions corrected for multiple hits is estimated by a likelihood approach, based on phylogeny and inferred ancestral sequences; and (2) the ML estimate of $\alpha$ is obtained under a negative binomial distribution (Uzzel and Corbin 1971) using the expected number of substitutions. Computer simulation has been conducted to examine the performance of the new method.

Method

Number of Substitutions at a Site

The number of substitutions at a site cannot be observed from the present-day sequences, so it has to be "inferred." The traditional method of inference invokes the parsimony principle (Fitch 1971), which tends to underestimate the true number of substitutions (see Nei...
called the number of changes. For given sequence data
branch is long, resulting in \( m \) I
provide a more efficient approach for estimating \( m \) than
principle (e.g., Schluter 1995; Yang, Kumar, and Nei
is important to distinguish between the number of sub-
amino acids (or nucleotides) at the two ends of a branch
possibility of multiple hits is completely neglected. Note
that recently developed methods for ancestral sequence
is noteworthy that
In the following, we will develop a new method to
For a given site, the branches along the tree can be
For simplicity, we only discuss amino acid se-
quencies; it is virtually the same for nucleotide sequenc-
sequences, whose phylogenetic tree (topology) is
known or can be inferred. It is known that the total
number of branches for an unrooted tree is \( M = 2n - 3 \), or \( M = 2n - 2 \) for a rooted tree. At a given site, we
assume that \( k \) along the tree follows a Poisson distri-
where \( B \) is the total branch length of the tree and \( u \) is
the evolutionary rate at this site. To avoid confusion, it
noteworthy that \( k \) under the Poisson model is a ran-
so it is not statistically meaningful to “es-
which will be used for estimating \( \alpha \).

The number of substitutions at the same site that
occur on branch \( i \) also follows a Poisson distribution
with the expectation \( u b_i \), where \( b_i \) is the length of branch
obviously, we have \( u b_i - k b_i / B \). Thus, the probability
of no change on branch \( i \) (i.e., the amino acids at the
two ends of this branch are the same) is given by
and the probability of a change (i.e., the amino acids are
different at the two ends of the branch) is
For a given site, the branches along the tree can be
divided into two groups. The first group, denoted by \( G_1 \),
includes the branches on which (amino acid) changes
occur, and the second group, denoted by \( G_0 \), includes
the branches on which no changes occur. Obviously, the
total number of branches in \( G_1 \) is equal to \( m \) at the site,
and the total number of branches on \( G_0 \) is therefore
given by \( M - m \). Then, when the information about
branches in \( G_1 \) and \( G_0 \) at a site is known, the (conditional)
likelihood function can be written as
\[
L = \prod_{i \in G_1} q_i \prod_{j \in G_0} q_j \prod_{j \in G_0} p_j
= \prod_{i \in G_1} (1 - e^{-k b_i / B}) \prod_{j \in G_0} e^{-k b_j / B}.
\]
Tajima and Nei's (1984) model, where \( c \) is the frequency of nucleotide \( i \). For a given site, the likelihood function in equation (4) can be now modified as follows:

\[
L = \prod_{i=1}^{m} c(1 - e^{-kh_i}) \prod_{i=m+1}^{M} (1 - c + ce^{-kh_i})
\]  

(11)

Similar to equation (7), \( \hat{k} \) at this site can be estimated by letting \( \theta \ln L / \hat{k} = 0 \), which yields:

\[
\sum_{i=1}^{m} \frac{b_i/B}{1 - e^{-kh_i}} + (1 - c) \sum_{i=m+1}^{M} \frac{b_i/B}{1 - c + ce^{-kh_i}} = c.
\]  

(12)

Clearly, equation (12) will be reduced to equation (7) when \( c = 1 \).

In the above formulation, we assume that ancestral sequences are known. Thus, under a given phylogenetic tree with estimated branch lengths, it is straightforward to classify a branch into \( G_1 \) or \( G_0 \) by simply comparing the amino acids at the two ends of the branch at each site. The number of changes along the tree (\( m \)) is therefore easily counted. Note that in Felsenstein (1981), ancestral sequences are integrated out since they are considered unobservable. The methods for the phylogenetic reconstruction and branch length estimation have been well developed (for a recent review, see Nei 1996). The ancestral sequences can be inferred by either parsimony- or likelihood-based methods (e.g., Eck and Dayhoff 1966; Libertini and Di Donato 1994; Schluter 1995; Yang, Kumar, and Nei 1995; Koshi and Goldstein 1996; Zhang and Nei 1997). In the current study, Zhang and Nei's (1997) algorithm is used. In this method, the amino acid assignment that has the highest (posterior) probability is chosen to represent the inferred ancestral amino acids at this site.

Note that when a particular branch is very short, the least-squares estimate of the branch length can be negative because of the sampling error. In this case, the branch length is assigned to be zero in our algorithm. Since \( (b_i/B)/1 - e^{-kh_i} = 1/k \) as \( b_i = 0 \), the bias caused by our treatment in the zero-boundary seems trivial.

Estimation of \( \alpha \)

If amino acid (or nucleotide) substitutions at each site follow a Poisson process, and the substitution rate \( \lambda \) varies among sites according to the gamma distribution described by equation (1), the number of sites with the occurrence of \( k \) substitutions follows a negative binomial distribution, i.e.,

\[
f(k) = \frac{\Gamma(\alpha + k)}{k! \Gamma(\alpha)} \left( \frac{D}{D + \alpha} \right)^{\alpha} \left( \frac{\alpha}{D + \alpha} \right)^{k},
\]  

(13)

(Johnson and Kotz 1969), where \( D \) is the average number of substitutions per site along the tree.

The ML approach for estimating the parameter \( \alpha \) from a negative binomial distribution was clearly discussed by Johnson and Kotz (1969); it was also used by Sullivan, Holsinger, and Simon (1995) and Tourasse and Gouy (1997) for the rate variation among sites. In our case, the difference from the standard algorithm is that the number of substitutions at a site is replaced by its expectation, which can be estimated by equation (7) or equation (12). Therefore, the log-likelihood function can be written as

\[
\ln L = \sum_{i=1}^{N} \ln f(\hat{k}_i),
\]  

(14)

where \( N \) is the total number of sites and \( \hat{k}_i \) is the estimate of the expected number of substitutions at site \( i \), which is not necessarily an integer. One can easily show that the ML estimate of \( D \) is the same as that for the normal case, which is given by \( \hat{D} = \frac{\sum \hat{k}_i}{N/N} \). There is no simple solution for the estimate of \( \alpha \), but it can be numerically obtained, and the sampling variance of \( \hat{\alpha} \) can also be approximately obtained.

Implementation and Data Analysis

For aligned amino acid sequences, sites containing gaps will be excluded from further study. The current version of our algorithm assumes that the phylogenetic tree is known or can be appropriately inferred. The branch lengths for the given tree are estimated by the ordinary least-squares method using the Poisson correction or the method of Ota and Nei (1994), and the ancestral sequences can be inferred by a fast algorithm developed by Zhang and Nei (1997), in which Jones, Taylor, and Thornton's (1992) empirical model is used for computing the (posterior) probability of each amino acid assignment at ancestral nodes for a given site. Then, we can estimate the expected number of substitutions for each site by equation (7) or equation (12), and we can estimate \( \alpha \) by numerically maximizing the likelihood function of equation (14). We have also developed a similar algorithm for DNA sequences.

When the gamma parameter \( \alpha \) is estimated, the evolutionary distance corrected for the rate variation among sites can be estimated (e.g., Jin and Nei 1990; Tamura and Nei 1993; Gu and Li 1996; Tourasse and Gouy 1997; Gu 1997). Some useful distance measures for both amino acid and DNA sequences are summarized in table 1. These distance measures can be used for phylogenetic reconstruction, constant-rate testing, and divergence time estimation. For example, the effect of among-site rate variation on the phylogenetic inference can be examined by using an iterative algorithm: (1) the first tree is inferred assuming a uniform rate among sites; and (2) the parameter \( \alpha \) is estimated by the current method based on the first tree, and then the phylogeny is inferred by using these improved distances. The effect of among-site rate variation is nontrivial if these two trees are significantly different. Additional iterations may be needed to reach the convergence.

Table 2 shows the estimates of \( \alpha \) of 13 proteins encoded by mammalian mitochondrial genomes by the
Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>DNA</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC</td>
<td>(d = 3\alpha(1 - p/3)^{1/6} - 1)</td>
<td>JC-protein</td>
</tr>
<tr>
<td>K2P</td>
<td>(d = 4\alpha(2p - 2)Q^{1/2} + (1 - 2Q)^{1/2b} - 1)</td>
<td>Gu97</td>
</tr>
<tr>
<td>TN</td>
<td>(d = 4\alpha(1 - p)^{1/6} - 1)</td>
<td>Poisson</td>
</tr>
<tr>
<td>SRV</td>
<td>(d = 2\alpha - 1)</td>
<td>Gu97</td>
</tr>
</tbody>
</table>

Methods:
- DNA: JC (Jukes and Cantor 1969); K2P (Kimura 1980), TN (Tajima and Nei 1984), and SRV (Gu and Li 1996).
- Protein: Poisson, JC-protein, and Gu97.

Parameters: \(\alpha\) is the gamma distribution parameter defined by equation (1); \(p\) is the proportion of nucleotide or amino acid differences between two sequences.

Table 2

<table>
<thead>
<tr>
<th>Genes</th>
<th>Moments</th>
<th>Sullivan</th>
<th>YK</th>
<th>New</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atp6</td>
<td>1.18</td>
<td>1.06</td>
<td>0.64</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Atp8</td>
<td>51.02</td>
<td>51.02</td>
<td>43.60</td>
<td>1.48</td>
<td>0.92</td>
</tr>
<tr>
<td>Co1</td>
<td>0.78</td>
<td>0.65</td>
<td>0.39</td>
<td>0.21</td>
<td>0.27</td>
</tr>
<tr>
<td>Co2</td>
<td>1.89</td>
<td>1.76</td>
<td>0.99</td>
<td>0.22</td>
<td>0.49</td>
</tr>
<tr>
<td>Co3</td>
<td>0.57</td>
<td>0.51</td>
<td>0.32</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>Cytb</td>
<td>1.35</td>
<td>0.94</td>
<td>0.55</td>
<td>0.43</td>
<td>0.36</td>
</tr>
<tr>
<td>Nd1</td>
<td>3.82</td>
<td>2.77</td>
<td>1.27</td>
<td>0.61</td>
<td>0.57</td>
</tr>
<tr>
<td>Nd2</td>
<td>22.41</td>
<td>20.75</td>
<td>3.35</td>
<td>1.31</td>
<td>0.90</td>
</tr>
<tr>
<td>Nd3</td>
<td>2.13</td>
<td>1.86</td>
<td>1.03</td>
<td>0.45</td>
<td>0.34</td>
</tr>
<tr>
<td>Nd4</td>
<td>2.50</td>
<td>2.06</td>
<td>1.11</td>
<td>0.62</td>
<td>0.57</td>
</tr>
<tr>
<td>Nd41</td>
<td>7.05</td>
<td>5.31</td>
<td>1.77</td>
<td>0.85</td>
<td>0.79</td>
</tr>
<tr>
<td>Nd5</td>
<td>3.15</td>
<td>2.37</td>
<td>1.18</td>
<td>0.78</td>
<td>0.61</td>
</tr>
<tr>
<td>Nd6</td>
<td>13.50</td>
<td>12.76</td>
<td>2.86</td>
<td>1.06</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Methods:
- Moments (e.g., Tamura and Nei 1993) and Sullivan (1994) method.
- YK is Yang and Kumar’s (1996) method.

Note: The protein sequences of the rat, mouse, cow, fin whale, and blue whale were used in the estimation: the (unrooted) phylogenetic tree was ((rat, mouse), (cow, fin whale, blue whale)).

The performance of our new method was compared to that of three previous methods and the ML method with discrete gamma approximation (Yang 1994). In the method of moments (e.g., Tamura and Nei 1993) and Sullivan, Holsinger, and Simon’s (1995) method, parsimony was used to infer the number of substitutions. The only difference between them is that \(\alpha\) is estimated by the method of moments or by a likelihood approach (Sullivan, Holsinger, and Simon 1995). The third method is Yang and Kumar’s (1996) method, discussed above.

First, we compare the performance of our method with that of the method of moments, Sullivan, Holsinger, and Simon’s (1995) method, and Yang and Kumar’s (1996) method (YK). For \(\alpha = 0.5\) (table 3), the performance of the new method is always better than that of these methods. For example, in the case of low divergence in tree A, the mean for 300 replications is \(\alpha = 0.56 \pm 0.23\) when the sequence length \(L = 100\), and \(\alpha\) new method and by other methods. For each protein, the amino acid sequences from the mouse, rat, cow, fin whale, and blue whale were used (see Russo, Takezaki, and Nei [1996] for the species names and sequences). The estimates by our method are similar to those by the likelihood method, whereas other methods generally give larger estimates. In this case, Yang and Kumar’s (1996) method tends to overestimate \(\alpha\) because some branches (rat/mouse, or fin/blue whales) are short, while the internal branches among the three mammalian orders are quite long. We also analyzed these proteins from 11 vertebrate species (see Russo, Takezaki, and Nei 1996), and a similar pattern was observed.

Computer Simulation

The statistical properties of the new method and some other methods have been examined by computer simulation. Two model trees, each with eight protein sequences, are used (fig. 1); one is symmetric (A) and the other is asymmetric (B). In our simulation study, amino acid substitutions are assumed to follow a Poisson model. We consider two sets of branch lengths for each model tree: one is of low divergence (i.e., the branch lengths are short) and the other is of high divergence (i.e., the branch lengths are long); see the legend of figure 1 for details. The sequence length in the simulation is 100 or 300 amino acids. For each case, 500 replications of simulations were conducted (100 replications for the ML method).

After eight present-day amino acid sequences are generated according to the model tree, the procedure of estimation is the same as described before (using eq. 7). In our simulation, the parameter \(\alpha\) is set to be 0.5, 1.0, and 2.0, respectively, which approximately represents strong, intermediate, and weak rate variation among sites (Gu, Fu, and Li 1995).

![Tree A](https://example.com/treeA.png)

![Tree B](https://example.com/treeB.png)

**FIG. 1.**—The model trees used for simulation study. Branch lengths: (1) in the case of low divergence, \(a = 0.02\), \(b = 0.03\), \(c = 0.05\), \(d = 0.01\), and \(e = 0.04\); and (2) in the case of high divergence, \(a = 0.08\), \(b = 0.12\), \(c = 0.20\), \(d = 0.04\), and \(e = 0.16\).
1110 Gu and Zhang

Table 3
Simulation Results in the Case of \( \alpha = 0.5 \)

<table>
<thead>
<tr>
<th>Method</th>
<th>LOW DIVERGENCE</th>
<th>HIGH DIVERGENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 a.a</td>
<td>300 a.a</td>
</tr>
<tr>
<td>Tree A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moments</td>
<td>0.92 ± 0.67</td>
<td>0.79 ± 0.20</td>
</tr>
<tr>
<td>Sullivan</td>
<td>0.85 ± 0.70</td>
<td>0.71 ± 0.20</td>
</tr>
<tr>
<td>YK</td>
<td>0.60 ± 0.40</td>
<td>0.53 ± 0.14</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>0.52 ± 0.34</td>
<td>0.51 ± 0.14</td>
</tr>
<tr>
<td>New</td>
<td>0.56 ± 0.23</td>
<td>0.51 ± 0.12</td>
</tr>
<tr>
<td>Tree B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moments</td>
<td>0.98 ± 0.70</td>
<td>0.86 ± 0.22</td>
</tr>
<tr>
<td>Sullivan</td>
<td>0.89 ± 0.68</td>
<td>0.76 ± 0.22</td>
</tr>
<tr>
<td>YK</td>
<td>0.64 ± 0.40</td>
<td>0.57 ± 0.16</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>0.53 ± 0.31</td>
<td>0.50 ± 0.15</td>
</tr>
<tr>
<td>New</td>
<td>0.51 ± 0.23</td>
<td>0.48 ± 0.11</td>
</tr>
</tbody>
</table>

NOTE.—See table 2 footnote for abbreviations of methods.

= 0.51 ± 0.12 when \( L = 300 \); the number after ± is the square root of the sampling variance of the estimate. When the divergence is high, our method is also better than other methods, although the estimation bias is generally larger than that in the case of low divergence. Indeed, when the branch lengths are long, the estimate of the expected number of substitutions \( (k) \) is subject not only to large sampling variance but also to the error of ancestral sequence inference (Zhang and Nei 1997). A similar pattern has been observed in the case of \( \alpha = 1.0 \) and \( \alpha = 2.0 \) (see tables 4 and 5, respectively).

Two parsimony methods, the method of moments (Tamura and Nei 1993) and Sullivan, Holsinger, and Simon’s (1995) method, tend to overestimate \( \alpha \) in all cases. Wakeley (1993) obtained a similar result in the case of short-term (within-population) evolution. As expected, the bias of these methods increases with increasing \( \alpha \), increasing divergence, or decreasing sequence length. Generally, the estimation bias is more serious in the case of tree B than in that of tree A, indicating that the number of substitutions in a long branch can be seriously underestimated by the parsimony method.

The performance of YK is much better than that of the two parsimony methods. However, as shown in tables 3–5, the estimation bias of YK is generally larger than that of our method. The performance difference between our method and YK is more significant in the case of tree B than in that of tree A. This result is not surprising, because in YK, equal branch length is assumed.

Remarkably, for all cases we have examined, the performance of our method is very close to that of the ML method (Yang 1994). When \( \alpha \) is small (e.g., 0.5), the estimation bias of the ML method is slightly smaller than that of our method. However, this is not always the case when \( \alpha \) is larger (1.0 or 2.0); indeed, in some cases, our method is slightly better than the ML method. Overall, the sampling variances of the estimate are quite similar between the two methods. Therefore, since our method requires much less computational time than the ML method, it is useful in large-sequence data analysis.

The total computational time required for our method includes that for least-squares estimation of branch length, ancestral sequence inference, and estimation of \( m, k, \) and \( \alpha \), among which ancestral sequence inference

Table 4
Simulation Results in the Case of \( \alpha = 1.0 \)

<table>
<thead>
<tr>
<th>Method</th>
<th>LOW DIVERGENCE</th>
<th>HIGH DIVERGENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 a.a</td>
<td>300 a.a</td>
</tr>
<tr>
<td>Tree A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moments</td>
<td>2.69 ± 4.59</td>
<td>1.76 ± 1.36</td>
</tr>
<tr>
<td>Sullivan</td>
<td>3.01 ± 8.22</td>
<td>1.65 ± 1.37</td>
</tr>
<tr>
<td>YK</td>
<td>2.05 ± 7.74</td>
<td>1.14 ± 0.54</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>1.32 ± 1.20</td>
<td>1.09 ± 0.55</td>
</tr>
<tr>
<td>New</td>
<td>1.13 ± 0.77</td>
<td>0.98 ± 0.33</td>
</tr>
<tr>
<td>Tree B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moments</td>
<td>3.13 ± 5.83</td>
<td>1.92 ± 0.98</td>
</tr>
<tr>
<td>Sullivan</td>
<td>3.61 ± 9.75</td>
<td>1.80 ± 0.98</td>
</tr>
<tr>
<td>YK</td>
<td>2.31 ± 7.33</td>
<td>1.24 ± 0.54</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>1.32 ± 1.16</td>
<td>1.04 ± 0.38</td>
</tr>
<tr>
<td>New</td>
<td>1.13 ± 1.64</td>
<td>0.88 ± 0.25</td>
</tr>
</tbody>
</table>

NOTE.—See table 2 footnote for abbreviations of methods.
by analyzing many independent genes. On the other
hand, we have also recognized that our computational
ability in sequence analysis has recently been signific-
antly improved so that some sophisticated algorithms
can be implemented in practice.

Our method can be easily extended to the case of
the invariant + gamma model (Gu, Fu, and Li 1995)
which takes invariant sites into account. Tournasse
developed an algorithm for estimating $\alpha$ and
the proportion of invariant sites $\theta$ based on an "in-
variant + truncated negative binomial" distribution.
Unfortunately, the estimation bias in Tournasse’s
method caused by parsimony can be efficiently
corrected by using our algorithm (e.g., eq. 7). Unfortu-
ately, as shown by Gu, Fu, and Li (1995) and Tournasse
and Gouy (1997), the estimates under the invariant +
gamma model are subject to large sampling variance
except for very large numbers of sequences. Because of
the importance of invariant sites in molecular evolution
(e.g., rDNA), further study seems necessary.

Our method relies on the accuracy of the ancestral
sequence inference. In our implementation, we use
Zhang and Nei’s (1997) algorithm rather than others
(e.g., Eck and Dayhoff 1966; Libertini and Di Donato
1994; Schluter 1995; Yang, Kumar, and Nei 1995; Koshi
and Goldstein 1996) because it is computationally fast
and robust against many factors such as substitution
model, the phylogenetic tree, and the rate variation
among sites. Our study provides a good example of the
importance of developing a fast algorithm from the
viewpoint of practice (Nei 1996). For example, in prin-
ciple, Zhang and Nei’s (1997) and Yang, Kumar, and
Nei’s (1995) methods are the same, and their perfor-
ances are almost identical. Although Yang, Kumar,
and Nei’s (1995) method is statistically nice because the
unknown parameters (e.g., branch lengths) are estimated

is the speed-limiting factor. When Zhang and Nei’s
(1997) algorithm is implemented, the CPU time required
by all steps is only a few seconds, even for more than
20 sequences. We compared the (relative) CPU time of
our method with that of Yang’s (1994) discrete gamma
method. For the model trees used for simulation, our
method is, on average for all cases, over 100 times faster
than Yang’s (1994) method. When more sequences are
used, the difference in speed between our method and
the ML method is expected to be much greater.

Discussion

In this paper, we have developed an efficient meth-
method for estimating rate variation among sites (the gamma
distribution parameter $\alpha$). There are several simple
methods available for estimating $\alpha$ (e.g., Uzzel and Cor-in 1971; Tamura and Nei 1993; Sullivan, Holsinger,
and Simon 1995; Tournasse and Gouy 1997). The dis-
tinction of the new method is that the parsimony prin-
ciple, which tends to underestimate the number of sub-
stitutions, is not used. Instead, a likelihood approach has
been implemented in the new method for estimating the
expected number of substitutions, which is used for es-
imating $\alpha$. Use of an efficient algorithm for inferring
the ancestral sequences (Zhang and Nei 1997) is also
important in this procedure. Computer simulation has
shown that our method is generally better than the par-
simony-based methods.

Since it is less problematic to infer $m$ by the par-
simony method, Yang and Kumar (1996) suggested that
the likelihood of the number of changes ($m$) be used to
estimate $\alpha$. However, in their method, one has to assume
that all the branches of the tree are of equal length;
otherwise, the search for the ML estimate of $\alpha$ is no
longer fast. Therefore, as shown by the example and
simulation results, Yang and Kumar’s (1996) method
could be biased in some cases. In this sense, our method
is better, because it does not have such a problem.

The ML methods developed by Yang (1993, 1994)
and Gu, Fu, and Li (1995) for estimating $\alpha$ have nice
statistical properties but require much computational
time. Indeed, failure to handle large data sets will limit

Table 5
Simulation Results in the Case of $\alpha = 2.0$

<table>
<thead>
<tr>
<th>Method</th>
<th>LOW DIVERGENCE</th>
<th>HIGH DIVERGENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 a.a</td>
<td>300 a.a</td>
</tr>
<tr>
<td>Tree A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moments</td>
<td>4.87 ± 7.86</td>
<td>5.13 ± 6.43</td>
</tr>
<tr>
<td>Sullivan</td>
<td>5.88 ± 8.22</td>
<td>5.22 ± 7.22</td>
</tr>
<tr>
<td>YK</td>
<td>3.28 ± 7.74</td>
<td>2.96 ± 5.10</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>2.20 ± 1.10</td>
<td>2.06 ± 0.44</td>
</tr>
<tr>
<td>New</td>
<td>2.53 ± 0.77</td>
<td>1.90 ± 0.98</td>
</tr>
<tr>
<td>Tree B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moments</td>
<td>6.75 ± 10.5</td>
<td>6.28 ± 8.97</td>
</tr>
<tr>
<td>Sullivan</td>
<td>8.92 ± 18.4</td>
<td>6.40 ± 10.1</td>
</tr>
<tr>
<td>YK</td>
<td>6.03 ± 16.1</td>
<td>3.44 ± 7.04</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>2.32 ± 3.17</td>
<td>2.34 ± 1.54</td>
</tr>
<tr>
<td>New</td>
<td>2.16 ± 3.67</td>
<td>1.64 ± 0.82</td>
</tr>
</tbody>
</table>

Note.—See table 2 footnote for abbreviations of methods.
by the ML method, it is not suitable for our purposes because it requires too much computational time.

When equation (7) or equation (12) is used to estimate the expected number of substitutions, the branch lengths along the tree should be estimated first. In our algorithm, the rate variation among sites is not considered in this step, so the branch length could be underestimated. However, our simulation study has shown that the estimation bias caused by this problem is non-trivial only when some very divergent sequences are involved and $\alpha$ is very small (data not shown). If this is the case, the bias can be corrected by using one or two iterations, which is an option in our program.

In our method, we estimate the expected number of substitutions ($k$) for each site when ancestral sequences are inferred. Nielsen (1997) studied how to estimate the substitution rate site-by-site under the framework of Felsenstein (1981). However, Nielsen’s estimation seems to be subject to large sampling variance and bias; at some sites, the rate can be estimated to be as large as $\alpha$. Our approach has no such problem. A modified likelihood approach was proposed by Nielsen (1997) by introducing a prior distribution of rate, but it needs to “empirically” estimate the prior distribution first.

Although a great effort has been made to determine how to estimate the gamma parameter $\alpha$, the biological meaning of the model for rate variation among sites has not been well investigated. Indeed, gamma distribution has been suggested to model the among-site rate variation mainly because it seems to fit the pattern of rate variation well and is also mathematically simple. Therefore, the gamma distribution model is designed for statistical approximation rather than for a biological entity. In addition, current models always assume that the substitution rate, although it varies among sites, is the same among various evolutionary lineages at a given site. This implies that the functional role of each amino acid residue should not be changed in evolution, i.e., there is no functional divergence. Thus, it may be problematic to apply this model for gene family evolution because of the possible functional divergence among paralogous genes. Another problem concerns the biological meaning of $\alpha$. Since $\alpha$ is usually interpreted as an index of the rate variation among sites, it is expected that, for a particular gene, $\alpha$ should be a constant during evolution. The methodology we have developed in this paper provides an efficient tool for investigating these important issues.

Program Availability

A computer program for estimating the gamma shape parameter $\alpha$ and evolutionary distances from amino acid or DNA sequences is available on request.

Acknowledgments

We are grateful to Masatoshi Nei and Sudhir Kumar for helpful comments. This work was supported by research grants from the NIH and NSF to M. Nei.

LITERATURE CITED

RUSSO, A. M., N. TAKEZAKI, and M. NEI. 1996. Efficiencies of different genes and different tree-building methods in


MARCY K. UYENOYAMA, reviewing editor

Accepted August 5, 1997