The Age of the Common Ancestor of Eukaryotes and Prokaryotes: Statistical Inferences

Xun Gu

Institute of Molecular Evolutionary Genetics, The Pennsylvania State University

In this paper, a simple distance measure was used to estimate the age (T) of the common ancestor of eukaryotes and prokaryotes which takes the rate variation among sites and the pattern of amino acid substitutions into account. Our new estimate of T based on Doolittle et al.'s data is about 2.5 billion years ago (Ga), with 95% confidence interval from 2.1 to 2.9 Ga. This result indicates (1) that Doolittle et al.'s estimate (~2.0 Ga) seems too recent, and (2) that the traditional view about the divergence time between eukaryotes and prokaryotes (T_0 = 3.5 Ga) can be rejected at the 0.1% significance level.

Introduction

Establishing a reliable divergence timetable for major living forms of life is one of the most important issues in evolutionary study. Since fossil records are always incomplete and many gaps remain unsolved, use of molecular data has been suggested by many authors based on the concept of molecular clock (e.g., Zuckerkandl and Pauling 1965; McLaughlin and Dayhoff 1972; Dayhoff, Schwartz, and Orcutt 1978; Hori and Osawa 1979; Doolittle et al 1996*a*; Hedges et al. 1996; Wray, Levinton, and Shapiro 1996). By using 57 different protein sequences, Doolittle et al. (1996*a*) estimated that the divergence time (T) between eukaryotes and prokaryotes is about 2.0 billion years ago (Ga). This conclusion is sharply in contrast to the traditional view (about 3.5 Ga) and, so, highly controversial (for discussions, see Doolittle et al. 1996*b*; Gogarten et al. 1996; Hasegawa and Fitch 1996; Miyamoto and Fitch 1996; Mooers and Redfield, 1996; Morell, 1996).

Time estimation from sequence data relies on the evolutionary distance that is expected to be linear with time t. The linear property may be problematic when some assumptions (e.g., a uniform rate among sites) used for estimation are violated (Nei 1987; Zharkikh 1994). Doolittle et al. (1996*a*) discussed the effect of rate variation among sites but did not take it into account in their time estimation. If the uniform-rate assumption does not hold, the distance would be underestimated, and the bias tends to increase as the divergence increases (Jin and Nei 1990; Wakeley 1993; Gu, Fu, and Li 1995). Consequently, the divergence time of prokaryotes and eukaryotes (T) can be underestimated by the internal calibration (e.g., some recent fossil records of vertebrates). Indeed, this problem has become a major issue for debate since the publication of Doolittle et al.'s (1996*a*) paper (Doolittle et al. 1996*b*; Gogarten et al. 1996; Hasegawa and Fitch 1996; Miyamoto and Fitch 1996).

Miyamoto amd Fitch (1996) proposed a distance measure under the "invariant+gamma" model for the rate variation among sites, which also considered the pattern of amino acid substitutions. By investigating some combinations of "reasonable" parameter sets, Miyamoto amd Fitch (1996) concluded that the divergence time (T) between eukaryotes and prokaryotes could be anywhere between 2.0 Ga and 3.5 Ga. Although their study indicated the importance of accurate estimation of these parameters, Miyamoto and Fitch (1996) cannot show statistically which T = 2.0 or T = 3.5 Ga) is more likely. The same problem (i.e., lack of statistical inference for the time estimation) also existed in the discussions among Doolittle et al. (1996*b*), Gogarten et al. (1996), and Hasegawa and Fitch (1996).

The purpose of our paper is to correct the systematic bias in time estimation caused by neglecting the rate variation among sites, and to conduct a statistical inference to test whether the estimate of time (the age the common ancestor of eukaryotes and prokaryotes) is significantly less than the traditional view. Doolittle et al.'s (1996*a*) data will be reanalyzed by estimating the rate variation among sites for each protein. A simple distance measure is used which takes the pattern of amino acid substitutions and the rate variation among sites into account. The importance of statistical inferences for solving the current controversy is emphasized, and the biological implications of our results are discussed.

Methods

The Data Set

The amino acid sequences of 57 proteins used in Doolittle et al. (1996*a*) were obtained from the network (anonymous ftp: juno.ucsd.edu) or from GenBank. For each protein, amino acid sequences were aligned by the computer program CLUSTALV (Higgines, Bleasby, and Fuchs 1992) and then adjusted manually if necessary. A complete list of accession numbers of the 57 proteins discussed will be provided upon request to the author (e-mail: xungu@ imeg.bio.psu.edu).

The Evolutionary Distance

The distance measure (d) used for time estimation should consider two factors: the rate variation among sites and the pattern of amino acid substitutions. Conventionally, some empirical models for amino acid sub-

Key words: gamma distribution, evolutionary distance, substitution patterns, rate variation, common ancestor.

Address for correspondence and reprints: Xun Gu, Institute of Molecular Evolutionary Genetics, 328 Mueller Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802. E-mail: xungu@imeg.bio.psu.edu.

Mol. Biol. Evol. 14(8):861–866. 1997

Table 1
Comparison Among Some Distance Measures: Dayhoff's Distance, the New Distance (eq. 4), Kimura (1983), and Ota and Nei (1994)

p	Dayhoff	Equation (4)	Kimura (eq. 10)	Ota and Nei
0.01	0.01	0.01	0.01	0.01
0.05	0.05	0.05	0.05	0.05
0.10	0.11	0.11	0.11	0.11
0.20	0.23	0.23	0.23	0.23
0.30	0.38	0.37	0.38	0.39
0.40	0.56	0.55	0.57	0.58
0.50	0.80	0.77	0.80	0.83
0.60	1.12	1.07	1.11	1.16
0.70	1.59	1.54	1.60	1.65

NOTE.—P is the proportion of differences.

stitutions (e.g., Dayhoff, Schwartz, and Orcutt 1978; Jones, Taylor, and Thornton 1992) were used to compute the distance. However, if the rate variation among sites is considered, the estimation problem is no longer simple and can be only solved numerically (unpublished data). Therefore, it is desirable to develop a simple method for estimating d.

Consider a simple model for the pattern of amino acid substitutions: (1) the number of amino acids that are functionally acceptable at a given site is v ($1 \leq v \leq 20$), and the substitution rate between any two amino acids in this group is λ; and (2) for other $20 - v$ amino acids, any substitution is deleterious and is virtually not allowed in evolution, i.e., the substitution rate is zero. Under this model, one can show that, after t time units, the probability of difference between two sequences at a site is given by

$$p = b(1 - e^{-2\lambda t}), \tag{1}$$

where the constant $b = (v - 1)/v$. If all types of amino acids are allowable, v is equal to the number of amino acids ($v = 20$) and so $b = 0.95$. In this special case, equation (1) is reduced to the well-known Poisson model corrected for the multiple hits. From equation (1), the evolutionary distance between two amino acid sequences can be estimated by

$$d = -b \ln(1 - p/b) \tag{2}$$

(Bulmer 1991; Miyamoto and Fitch 1996). For instance, the Poisson distance corrected for the multiple hits ($b = 0.95$) is therefore given by

$$d = -0.95 \ln(1 - p/0.95). \tag{3}$$

An unsolved problem is how to determine the constant b to account for the pattern of amino acid substitutions. A simple way is to compare the performance of equation (2) to that of Dayhoff's distance. This "performance-equivalent approach" has been used by Kimura (1983), Bulmer (1991) and Ota and Nei (1994). We found that the distance computed by equation (3) with $b = 0.83$ gives the best fit to the Dayhoff's distance (table 1). Thus, the distance measure

$$d = -0.83 \ln(1 - p/0.83) \tag{4}$$

holds that the number of acceptable amino acids of a site is on average equal to 6 ($v \approx 5.88$), which is (approximately) equivalent to the Dayhoff's distance. We also compared equation (4) to Kimura's (1983) formula and Ota and Nei's (1994) formula. Although all of them are generally very close to Dayhoff's distance, Kimura's (1983) formula is slightly better than others when the divergence is large. The advantage of the new distance measure (eq. 4) is that it can be easily extended to the case when the rate varies among sites (see below).

When the substitution rate (λ) varies among sites according to a gamma distribution

$$\phi(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \tag{5}$$

the new distance measure can be modified as follows. First we note that the gamma distribution has two parameters: the shape parameter α describes the degree of rate variation among sites, and β is a scalar. The shape parameter α is inversely related to the degree of rate variation among sites. Roughly speaking, $\alpha = 0.5$, 1.0, and 2.0 correspond to strong, intermediate, and weak rate variation among sites, respectively, and $\alpha = \infty$ means a uniform rate. This model was first proposed by Uzzel and Corbin (1971) and has been used widely (e.g., Nei, Chakraborty, and Fuerst 1976; Holmquist et al. 1983; Jin and Nei 1990; Tamura and Nei 1993; Yang 1993; Gu, Fu, and Li 1995; Gu and Li 1996).

Under this model, the proportion of sites that are different between two sequences is given by

$$\begin{aligned} p &= b - b\int_0^{\infty} e^{-2\lambda t}\phi(\lambda)\,d\lambda \\ &= b - b(1 + 2\bar{\lambda}t/\alpha)^{-\alpha}, \end{aligned} \tag{6}$$

where $\bar{\lambda} = \alpha/\beta$ is the mean substitution rate (Jin and Nei 1990; Gu, Fu, and Li 1995). Thus, similar to the derivation of equation (2), the evolutionary distance, $d = 2b\bar{\lambda}t$, is given by

$$d = b\alpha\{(1 - p/b)^{-1/\alpha} - 1\}. \tag{7}$$

Note that if the substitution rate is uniform among sites, i.e., $\alpha \rightarrow \infty$, equation (7) is reduced to equation (2). Thus, analogous to equation (3), the gamma distance when $b = 0.95$ is given by

$$d = 0.95\alpha\{(1 - p/0.95)^{-1/\alpha} - 1\} \tag{8}$$

(Rzhestky and Nei 1994), and analogous to equation (4), the gamma distance accounting for the pattern of amino acid substitutions is given by

$$d = 0.83\alpha\{(1 - p/0.83)^{-1/\alpha} - 1\}. \tag{9}$$

For comparison, Kimura's (1983) distance,

$$d = -\ln(1 - p - 0.2p^2), \tag{10}$$

is also used in our study.

There are many methods proposed for estimating α (e.g., Uzzel and Corbin 1971; Tamura and Nei 1993; Yang 1993; Gu, Fu, and Li 1995; Sullivan, Holsinger, and Simon 1995). Gu (1996) developed a fast method for estimating α which does not require any knowledge of phylogeny. Since Gu's (1996) method is a simplified likelihood approach, the estimate of α is asymptotically unbiased. This method has two steps: the ML maximum-likelihood (ML) estimate of α is obtained from each three-sequence set, and the final estimate is the average over all possible combinations. It was shown that the asymptotic bias of the estimate α can be significantly reduced if an appropriate weight function is chosen for combining the estimates of three-sequence sets (Gu 1996). The computer program, which was originally developed for nucleotide sequences, has been modified for protein sequences under Dayhoff's model.

Time Estimation

Recent fossil records, e.g., those of vertebrates, are chosen for calibration because they are reasonably reliable and less controversial. In order to estimate the divergence time of speciation accurately, it is important to use many independent genes (Hedges et al. 1996). Therefore, in our study we use the mammal–mammal (t_1 = 100 MYA) and the mammal–bird (t_2 = 300 MYA) divergence times because they are available for 43 and 12 proteins, respectively. Some fossil records, e.g., amniote–amphibian and tetrapod–fish, will not be used, because the numbers of genes are only 4 and 5, respectively, in Doolittle et al.'s (1996*a*) data. It is difficult to obtain a reliable estimate for the divergence time between eukaryotes and archaebacteria, because the total number of genes available is too low.

Time estimation requires the assumption of rate constancy among lineages. In Doolittle et al.'s (1996*a*) data, molecular clock does not hold for some proteins. A better approach is to exclude proteins from further study if a constant rate was rejected significantly. Many methods have been proposed for this purpose (e.g., Wu and Li 1985; Gu and Li 1992; Takezaki, Rzhetsky, and Nei 1995). In our study, a gene will be excluded from further study if it is rejected by the rate-constant test (Gu and Li 1992).

In the following, the divergence time to estimate is always denoted by *T*, and the divergence time for calibration is denoted by *t*. For protein *i*, let $D_i = 2\lambda T$ be the average distance for which the divergence time is to be estimated, and let $d_i = 2\lambda t$ be the average distance of calibration. Thus, from this protein, *T* can be estimated by $\hat{T}_i = (D_i/d_i)t$.

Since time estimation from a single gene is usually subject to large sampling variance and could be highly biased, one should consider how to minimize the estimation error when multiple genes are available. For example, our preliminary result has shown that when t_1 (mammal–mammal) was used for the reference time, a simple (arithmetic) average of *T* among genes was subject to large sampling variance. A simple method is suggested as follows.

For protein *i*, let the distance ratio be $\rho_i = d_i/D_i$. Our goal is to find the estimate of *T* which minimizes the sum of squares (*J*)

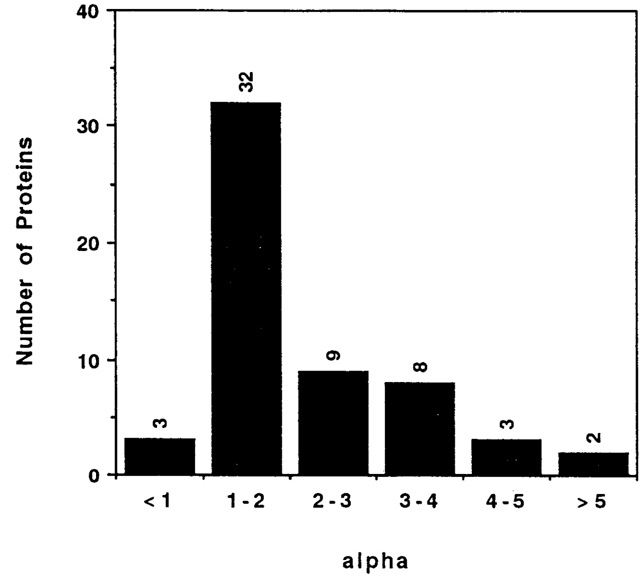

FIG. 1.—The histogram of estimates of α for 57 protein sequences.

$$J = \sum_{i=1}^{n} \left(\rho_i - \frac{t}{T}\right)^2, \qquad (11)$$

where *n* is the number of proteins used for the estimation. By $\partial J/\partial T = 0$, one can show that the estimate of time *T* is given by

$$\hat{T} = \frac{t}{\bar{\rho}}, \qquad (12)$$

where $\bar{\rho}$ is the average ratio

$$\bar{\rho} = \sum_{i=1}^{n} \rho_i/n. \qquad (13)$$

The large sampling variance of $\hat{T}$ can be approximately computed by

$$V(\hat{T}) = \left(\frac{\hat{T}}{\bar{\rho}}\right)^2 \mathrm{Var}(\bar{\rho}), \qquad (14)$$

where Var($\bar{\rho}$) is the sampling variance of $\bar{\rho}$.

Results

The gamma shape parameter α was estimated for all 57 protein sequences. Figure 1 shows that the estimates of α for most proteins are less than 2.0 (also see Holmquist et al. 1983; Russo, Takezaki, and Nei 1996). For a selected data set, we have estimated α by the ML method (Yang 1993; Gu, Fu, and Li 1995), and the maximum-parsimony (MP) method (Tamura and Nei 1993). Generally, our estimates are larger than ML estimates but considerably smaller than MP estimates (results not shown).

We denote the estimate of the age (*T*) of the common ancestor of eukaryotes and prokaryotes calibrated by the mammal–mammal clock by $\hat{T}_1$, and that calibrated by the mammal–bird clock by $\hat{T}_2$. As shown in table 2, when the new distance measure (eq. 9) is used for estimating *d* and *D* (i.e., the pattern of amino acid substitu-

Table 2
The Final Estimate of the Age ($\hat{T}$, Ga billions of years ago) of the Common Ancestor of Eukaryotes and Prokaryotes

Method	Substitution Pattern	Rate Variation	$\bar{\rho}$ (%)	$\hat{T}$
(1) Calibrated by the Mammal–Mammal Clock (t = 100 MYA, n = 38)				
New distance (eq. 9)	Yes	Yes	5.26 ± 0.70	2.43 ± 0.25
Rzhetsky and Nei (eq. 8)	No	Yes	5.52 ± 0.50	1.81 ± 0.16
New distance (eq. 4)	Yes	No	6.64 ± 0.67	1.51 ± 0.15
Kimura (eq. 10)	Yes	No	6.64 ± 0.67	1.51 ± 0.15
Poisson (eq. 3)	No	No	7.67 ± 0.77	1.30 ± 0.13
(2) Calibrated by the Mammal–Bird Clock (t = 300 MYA, n = 11)				
New method (eq. 9)	Yes	Yes	11.5 ± 1.40	2.61 ± 0.31
Rzhetsky and Nei (eq. 8)	No	Yes	13.6 ± 1.48	2.21 ± 0.24
New distance (eq. 4)	Yes	No	15.9 ± 1.78	1.89 ± 0.21
Kimura (eq. 10)	Yes	No	15.8 ± 1.78	1.90 ± 0.21
Poisson (eq. 3)	No	No	17.2 ± 1.76	17.4 ± 0.18

NOTE.—n is the number of proteins used.

tion and the rate variation among sites are considered), $\hat{T}_1$ is about 2.43 Ga and $\hat{T}_2$ is about 2.61 Ga. The 95% confidence interval of T_1 is between 1.94 Ga and 2.92 Ga, and the 95% confidence interval of T_2 is between 2.00 Ga and 3.22 Ga. On average, the estimate of T is 2.52 Ga, with a 95% confidence interval between 2.12 Ga and 2.92 Ga. In spite of the broad confidence interval, our result indicates that (1) the estimate of Doolittle et al. (1996*a*) (2.0 Ga) may be too recent, and (2) the origin of eukaryotes may not be much earlier than 3.0 Ga. Tables 3 and 4 show the time estimation from each gene based on the mammal–mammal clock and the mammal–bird clock, respectively. The results clearly indicate that the estimate of T based on a single gene, which is in a range from 1.2 Ga to much more than 4.5 Ga, is highly unreliable. Therefore, many genes should be used in order to obtain a reliable estimate of T.

Table 2 shows that the divergence time can be underestimated considerably if the rate variation among sites and/or the pattern of amino acid substitutions is neglected. For example, if the rate variation among sites is neglected, that is, equation (4) is used, $\hat{T}_1$ = 1.5 Ga and $\hat{T}_2$ = 1.9 Ga, which are 38% and 27% lower than the estimate given by the new distance measure (eq. 9), respectively. The estimate of T based on equation (10), the Kimura (1983) formula, is very similar to that based on equation (4). On the other hand, if the pattern of amino acid substitutions is not considered (i.e., eq. (8) is used), we have $\hat{T}_1$ = 1.8 Ga and $\hat{T}_2$ = 2.2 Ga, which are 26% and 16% lower than the estimate given by equation (9), respectively. The underestimation could be worse if both the rate variation among sites and the pattern of amino acid substitutions are neglected: $\hat{T}_1$ = 1.3 Ga and $\hat{T}_2$ = 1.7 Ga, which are 47% and 33% lower, respectively. Interestingly, time estimation calibrated by the mammal–bird clock ($\hat{T}_2$) seems more robust against the systematic bias of the distance measure than that calibrated by the mammal–mammal clock ($\hat{T}_1$).

Our understanding about the age of the common ancestor of eukaryotes and prokaryotes can be improved

Table 3
Estimates of the Age ($\hat{T}_1$) of Common Ancestor of Eukaryotes and Prokaryotes, Based on the Mammal–Mammal Calibration

Gene	α	$\rho = d/D$ (%)	$\phi\hat{T}_1$ (MYA)
Porphobilinogen synthase	3.49	8.33	1,200
Uroporphyrinogen decarboxylase	3.72	8.33	1,200
Thymidylate synthase	1.24	8.13	1,230
Argininosuccinate lyase	1.12	8.10	1,235
Acetyl CoA C-acetyl transferase	1.80	7.87	1,271
Cu/Zn SOD	1.32	7.59	1,318
Fructose bisphosphatase	1.91	7.34	1,362
Phosphoenolpyruvate carboxykinase	1.04	7.23	1,383
Nucleoside diphosphate kinase	1.68	7.04	1,421
Dihydrofolate reductase	1.99	6.72	1,488
Aspartate transaminase	2.01	6.35	1,575
Aldehyde dehydrogenase	1.73	6.17	1,612
Cysteine S-methyl transferase	0.66	5.28	1,893
Glyceraldehyde 3-phosphate dehydrogenase	1.76	4.99	2,004
Pyruvate kinase	1.69	4.90	2,041
Enolase	1.25	4.68	2,137
Ornithine carbamoyl transferase	3.27	4.51	2,217
L-Lactate dehydrogenase	3.77	4.44	2,252
Dihydrolipoamide S-acetyl transferase	3.26	3.97	2,519
Trypsin	1.66	3.72	2,688
G6PD	1.37	3.55	2,817
Alkaline phosphatase	1.51	3.17	3,155
Dimethylallyl transferase	1.82	3.06	3,268
Histidine-tRNA ligase	4.40	3.05	3,279
Asparagine synthase	1.45	2.79	3,584
TPI	1.34	2.75	3,636
Orotidine phosphate decarboxylase	2.09	2.20	4,545
Phosphoglycerate kinase	1.65	2.00	5,000
Dihydrolipoamide dehydrogenase	2.17	1.95	5,128
Phosphofructokinase	1.54	1.83	5,464
Glutamine-ammonia ligase	1.88	1.54	6,494
Peptidyl prolyl isomerase	0.98	1.52	6,579
Argininosuccinate synthase	2.64	1.39	7,194
DNA topoisomerase	3.29	1.01	9,901
Ribonucleotide reductase (large)	2.86	0.99	10,100
Ribonucleotide reductase (small)	1.56	0.98	10,204
Pyruvate dehydrogenase	1.64	0.87	11,494
HMG-CoA reductase	1.20	0.64	15,625

NOTE.—Distances are estimated by the new distance measure (eq. 9), and the divergence time is calibrated by the mammal–mammal fossil record (100 MYA).

Table 4
Estimate of the Age ($\hat{T}_2$) of the Common Ancestor of Eukaryotes and Prokaryotes, Based on the Mammal–Bird Calibration

Protein	α	ρ (%)	$\hat{T}_2$ (MYA)
Glycine dehydrogenase	0.99	18.59	1,612
Enolase	1.25	17.33	1,731
Aspartate transminase	2.01	15.95	1,881
Phosphoenolpyruvate carboxykinase	2.09	13.97	2,148
Dehydrofolate reductase	1.99	11.96	2,508
Aldehyde dehydrogenase	1.73	11.01	2,724
Pyruvate kinase	1.69	10.96	2,737
Lactate dehydrogenase	3.77	10.25	2,925
G3PD	1.76	7.33	4,095
Triose phosphate isomerase	1.31	6.91	4,441
Glutamate-ammonia ligase	1.88	3.05	9,800

NOTE.—Distances are estimated by the new distance measure (eq. 9), and the divergence time is calibrated by the mammal–bird fossil record (300 MYA).

if more fossil records are used. For example, the divergence time between chordate–arthropoda may be useful because many genes are available. Furthermore, the extrapolation based on them is expected to be more reliable than that based on the mammal–mammal clock. However, the time estimation based on these calibrations could be complicated by the fact that the divergence times of animal phyla are still highly controversial (e.g., Wray, Levinton, and Shapiro 1996). Based on the vertebrate calibration (mammal–mammal and/or mammal–bird), the date of the Drosophila–Vertebrate split is estimated to be about 800 MYA from an analysis of 17 proteins, which is about 10% earlier than the estimate of Doolittle et al. (1996*a*) (~700 MYA). We estimated the mean ratio of d to D over these proteins as $\bar{\rho} = 0.264$. Thus, if the split time of Drosophila–Vertebrate is about 700–800 mya, the age of the common ancestor of prokaryotes and eukaryotes can be estimated to be $T = 2.7$–3.0 Ga, which is consistent with the estimates based on the vertebrate calibration.

Discussion

For estimating the divergence times from protein sequence data, it is important to compute the evolutionary distance (d) accurately. In this paper, a simple distance measure was used, which takes the rate variation among sites and the pattern of amino acid substitutions into account. We reanalyzed Doolittle et al.'s (1996*a*) data and found that the age (T) of the common ancestor of eukaryotes and prokaryotes is about 2.5 Ga, with a 95% confidence interval from 2.1 Ga to 2.9 Ga. The current estimate is about 0.5 Ga earlier than that of Doolittle et al. (1996*a*). It seems that the difference is mainly due to the fact that the rate variation among sites was underestimated by Doolittle et al. (1996*a*).

During the revision of this manuscript, two technical comments as well as Doolittle et al.'s response appeared in *Science*; all of them were concerned with the problem of rate variation among sites. Hasegawa and Fitch (1996) pointed out that the proportion of invariable sites assumed by Doolittle et al. (1996*a*) (~5%) is too small. Doolittle et al.'s (1996*a*) work was also criticized by Gogarten et al. (1996) for two reasons: (1) horizontal gene transfer was not properly considered, and (2) the effect of rate variation among amino acid sites was significantly underestimated. By using $\alpha = 0.7$, Gogarten et al. (1996) contended that the date of prokaryote–eukaryote divergence can be more than 3.5 billion years ago. In their response, Doolittle et al. (1996*b*) argued that $\alpha = 0.7$ was too extreme, and $\alpha = 1.0$ would be more appropriate. In this regard, Doolittle et al. (1996*b*) modified their previous result and reassured an adjusted divergence time of 2.5 billion years ago. Miyamoto and Fitch (1996) also addressed this problem using a similar approach under a more complex model.

These discussions are helpful to understand how the rate variation among sites affects the time estimation, but the current study has two advantages over them: First, for each protein, the gamma parameter α was estimated from sequence data. Otherwise, one could not solve the debate between Gogarten et al. (1996) and Doolittle et al. (1996*b*). Second, the 95% confidence interval of our time estimate was given. Without statistical inferences, it is difficult to test whether the age of the common ancestor of eukaryotes and prokaryotes (T) estimated by protein sequences is significantly less than the traditional view ($T_0 = 3.5$ Ga).

We have recognized that our estimate of T is still quite rough. The first reason involves the methodology: When the divergence is large (say, $d > 1$), the estimate of d may be subject to large sampling variance and very sensitive to the estimate of α. Unfortunately, the estimation of α generally suffers from large sampling variance (Gu, Fu, and Li 1995). Therefore, the sampling variance of d can be considerably inflated when the sequence is short and the divergence is large. There is no doubt that much theoretical work needs to be done. The second reason involves the sequence data. For example, if the horizontal gene transfer had occurred between eukaryotes and prokaryotes (actually, it did in some genes we used), the divergence time would be underestimated. On the other hand, if some paralogous rather than orthologous genes are used, the divergence time would be overestimated. Therefore, a phylogeny-based approach is suggested to exclude sequences with these problems. More extensive and careful analysis can be done as many microbial genome sequences come out in the near future.

These caveats notwithstanding, we attempt to explore the biological significance of our statistical inferences. Paleobiologists have found 3.5-billion-year-old microfossils that look like extant cyanobacteria. If living organisms existed as much as 3.5 Ga, the key problem generated by Doolittle et al. (1996*a*) is whether eukaryotes appeared shortly after the origin of cellular life or took a long time to evolve. To answer this question, statistical inference is important. Taking the traditional view ($T_0 = 3.5$ Ga) as the null hypothesis, one can test whether the estimate of the last common ancestor of prokaryotes and eukaryotes ($\hat{T}$) from molecular data is significantly less than T_0. The null hypothesis (the traditional view) cannot be rejected by the molecular data when the upbound of the 95% confidence interval of $\hat{T}$ exceeds $T_0 = 3.5$ Ga. In our case, we obtained $\hat{T} = 2.5 \pm 0.2$ Ga. Therefore, the traditional view can be rejected at the 0.1% significance level (t-test). On the other hand, at the 5% significance level, we can conclude that the age of the last common ancestor of prokaryotes and eukaryotes is unlikely to be younger than 2 Ga.

Acknowledgments

The author thanks Dr. Masatoshi Nei and Dr. Sudhir Kumar for discussions and three anonymous reviewers for helpful comments. The author is particularly grateful to the reviewing editor, Dr. Naoyuki Takahata, for his invaluable help. This work was partially supported by NSF and NIH grants to M. Nei.

LITERATURE CITED

BULMER, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8**:868–883.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary changes in protein S. Pp. 345–352 *in* Atlas of protein sequence and structure. Vol. 5, Suppl. 3. Natural Biomedical Research Foundation, Washington, D.C.

DOOLITTLE, R. F., D. F. FENG, S. TSANG, G. CHO, and E. LITTLE. 1996*a*. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science **271**:470–477.

———. 1996*b*. Dating the Cenancester of organisms. Science **274**:1751-1753.

GOGARTEN, J. P., L. OLENDZENSKI, E. HILARIO, C. SIMON, and K. E. HOLSINGER. 1996. Dating the Cenancester of organisms. Science **274**:1751–1753.

GU, X. 1996. Models of DNA sequence evolution. Ph.D dissertation. Graduate School of Biomedical Sciences, University of Texas at Houston.

GU, X., Y. X. FU, and W. H. LI. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. **12**:546–557.

GU, X., and W. H. LI. 1992. Higher rates of amino acid substitution in rodents than in human. Mol. Phylogenet. Evol. **1**:211–214.

GU, X., and W. H. LI. 1996. A general additive distance with time-reversibility and rate variation among nucleotide sites. Proc. Natl. Acad. Sci. USA **93**:4671–4676.

HASEGAWA, M., and W. M. FITCH. 1996. Dating the Cenancester of organisms. Science **274**:1751–1753.

HEDGES, S. B., P. H. PARKER, C. G. SIBLEY, and S. KUMAR. 1996. Continental breakup and the ordinal diversification of birds and mammals. Nature **381**:226–229.

HIGGINES, D. G., A. J. BLEASBY, and R. FUCHS. 1992. CLUSTALV: improved software for multiple sequence alignment. Comput. Appl. Biosci. Comm. **8**:189–191.

HOLMQUIST, R. M., M. GOODMAN, T. CONROY, and J. CZELUSNIAK. 1983. The spatial distribution of fixed mutations within genes coding for proteins. J. Mol. Evol. **19**:437–448.

HORI, H., and S. OSAWA. 1979. Evolutionary changes in 5S RNA secondary structure and a phylogenetic tree of 54 5S RNA species. Proc. Natl. Acad. Sci. USA **76**:381–385.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony model of phylogenetic analysis. Mol. Biol. Evol. **7**: 82–102.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

MCLAUGHLIN, P. J., and M. O. DAYHOFF. 1972. Evolution of species and proteins: a time scale. Pp. 47–52 *in* Atlas of protein sequence and structure. Natural Biomedical Research Foundation, Washington, D.C.

MIYAMOTO, M. M., and W. M. FITCH. 1996. Constraints on protein evolution and the age of the eubacteria/eukaryote split. Syst. Biol. **45**:568-575.

MOOERS, A., and R. J. REDFIELD. 1996. Digging up the roots of life. Nature **379**:587–588.

MORELL, V. 1996. Proteins 'clock' the origins of all creatures—great and small. Science **271**:448.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

NEI, M., R. CHAKRABORTY, and P. A. FUERST. 1976. Infinite allele model with varying mutation rate. Proc. Natl. Acad. Sci. USA **73**:4164–4168.

OTA, T., and M. NEI. 1994. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J. Mol. Evol. **38**:642-643.

RUSSO, C. A., M. N. TAKEZAKI, and M. NEI. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. **13**:525–536.

RZHETSKY, A., and M. NEI. 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J. Mol. Evol. **38**:295–299.

SULLIVAN, J. K., K. E. HOLSINGER, and C. SIMON. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmontine rodents. Mol. Biol. Evol. **12**:988–1001.

TAKEZAKI, N., A. RZHETSKY, and M. NEI. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. **12**:823–833.

TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control of mitochondrial DNA in human and chimpanzees. Mol. Biol. Evol. **10**:512–526.

UZZEL, T., and K. W. CORBIN. 1971. Fitting discrete probability distribution to evolutionary events. Science **172**:1089–1096.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37**:613–623.

WRAY, G. A., J. S. LEVINTON, and L. H. SHAPIRO. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. Science **274**:568–573.

WU, C. I., and W. H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

YANG, Z., and S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13**: 650–659.

ZHARKIKH, A. 1994. Comparison of methods of estimation of evolutionary distance between nucleotide sequences. J. Mol. Evol. **39**:315–329.

ZUCKERKANDL, E., and L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–166 *in* V. BRYSON and H. J. VOGEL, eds. Evolving genes and proteins. Academic Press, New York.

NAOYUKI TAKAHATA, reviewing editor

Accepted April 29, 1997