

Correlation Between the Substitution Rate and Rate Variation Among Sites in Protein Evolution

Jianzhi Zhang and Xun Gu

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received December 10, 1997

Accepted for publication April 7, 1998

ABSTRACT

It is well known that the rate of amino acid substitution varies among different proteins and among different sites of a protein. It is, however, unclear whether the extent of rate variation among sites of a protein and the mean substitution rate of the protein are correlated. We used two approaches to analyze orthologous protein sequences of 51 nuclear genes of vertebrates and 13 mitochondrial genes of mammals. In the first approach, no assumptions of the distribution of the rate variation among sites were made, and in the second approach, the gamma distribution was assumed. Through both approaches, we found a negative correlation between the extent of among-site rate variation and the average substitution rate of a protein. That is, slowly evolving proteins tend to have a high level of rate variation among sites, and vice versa. We found this observation consistent with a simple model of the neutral theory where most sites are either invariable or neutral. We conclude that the correlation is a general feature of protein evolution and discuss its implications in statistical tests of positive Darwinian selection and molecular time estimation of deep divergences.

THE rate of amino acid substitution varies considerably among different proteins as well as different sites of a protein (*e.g.*, Kimura 1983; Nei 1987; Li 1997). It is interesting to study the relationship between these two types of rate variations because they result from the variation of functional constraints at different levels (proteins or amino acid residues) according to the neutral theory of molecular evolution (Kimura and Ohta 1974; Kimura 1983). In a recent study of vertebrate mitochondrial genes, Kumar (1996) observed that the extent of rate variation among nucleotide sites of a gene and the average substitution rate of the gene are negatively correlated. That is, fast evolving genes have low degrees of among-site rate variation, whereas slowly evolving genes have high degrees of among-site rate variation. However, the cause of the correlation was unclear at the time. Furthermore, the result was obtained based on the strong assumption that the among-site rate variation follows a gamma distribution that may or may not hold in all the genes. To examine whether such correlation is a general feature of molecular evolution, a large-scale analysis of nuclear and mitochondrial genes with an approach that is free from the assumption of the gamma distribution seems to be necessary. In this article, we analyzed protein sequences for 51 nuclear genes of vertebrates and 13 mitochondrial genes of mammals and found that the correlation holds for both nuclear and mitochondrial genes. We also explain the

underlying cause of the correlation and discuss its implications.

DATA AND METHODS

Characterization of among-site rate variation: Let us assume that the substitution rate (r) at an amino acid site of a protein sequence is an identically and independently distributed random number following an unknown distribution $g(r)$. Let k be the number of substitutions at a given site during a period of evolutionary time T . Assuming that amino acid substitutions follow the Poisson process when the rate is given, k follows the distribution:

$$f(k) = \int_0^{\infty} \frac{(rT)^k e^{-rT}}{k!} g(r) dr. \quad (1)$$

The expectation of k , $E(k)$, can be obtained by

$$\begin{aligned} E(k) &= \sum_{k=0}^{\infty} [kf(k)] \\ &= \sum_{k=0}^{\infty} \left[k \int_0^{\infty} \frac{(rT)^k e^{-rT}}{k!} g(r) dr \right] \\ &= \int_0^{\infty} \sum_0^{\infty} \left[k \frac{(rT)^k e^{-rT}}{k!} \right] g(r) dr \\ &= \int_0^{\infty} rTg(r) dr \\ &= \mu T, \end{aligned} \quad (2)$$

where $\mu = E(r)$ is the expectation of r , or mean substitution rate. The variance of k , $\text{Var}(k)$, can be obtained by

Corresponding author: Jianzhi Zhang, 322 Mueller Laboratory, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802. E-mail: jxz128@psu.edu

$$\begin{aligned}
\text{Var}(k) &= \sum_{k=0}^{\infty} [k^2 f(k)] - [E(k)]^2 \\
&= \sum_{k=0}^{\infty} \left[k^2 \int_0^{\infty} \frac{(rT)^k e^{-rT}}{k!} g(r) dr \right] - (\mu T)^2 \\
&= \int_0^{\infty} \sum_0^{\infty} \left[k^2 \frac{(rT)^k e^{-rT}}{k!} \right] g(r) dr - (\mu T)^2 \\
&= \int_0^{\infty} (r^2 T^2 + rT) g(r) dr - (\mu T)^2 \\
&= \text{var}(r) T^2 + (\mu T)^2 + \mu T - (\mu T)^2 \\
&= \text{Var}(r) T^2 + \mu T,
\end{aligned} \tag{3}$$

where $\text{Var}(r)$ is the variance of r . Therefore, the mean and variance of r can be obtained from

$$\mu = E(r) = E(k)/T \tag{4}$$

and

$$\text{Var}(r) = [\text{Var}(k) - E(k)]/T^2. \tag{5}$$

The degree of among-site rate variation is better measured by the coefficient of variation, which is obtained from

$$CV(r) = \frac{\sqrt{\text{Var}(r)}}{E(r)} = \frac{\sqrt{\text{Var}(k) - E(k)}}{E(k)}. \tag{6}$$

It is interesting to see from Equations 4 and 6 that if we can estimate the mean and variance of k , we will be able to study the mean substitution rate and the extent of among-site rate variation without knowing the underlying rate distribution $g(r)$ (Gu and Li 1998).

Nevertheless, the gamma distribution has been predominantly used in recent years to characterize the among-site rate variation of proteins (*e.g.*, Uzzell and Corbin 1971; Golding 1983; Tamura and Nei 1993; Yang 1994; Gu *et al.* 1995; Gu and Zhang 1997). That is, the substitution rate (r) at a site is assumed to follow the gamma distribution with the probability density function

$$g(r) = \frac{(\alpha/\mu)^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-(\alpha/\mu)r}, \tag{7}$$

where $\Gamma(\)$ is the gamma function. There are two parameters in Equation 7: the mean rate $\mu = E(r)$ and the shape parameter α . The extent of among-site rate variation can be measured by α . Small values of α indicate a high degree of rate variation among sites, whereas large values indicate a low degree of variation. In fact, under the gamma model, $CV(r) = 1/\sqrt{\alpha}$. Some examples of the gamma distribution with different values of μ and α are given in Figure 1.

Genes used: We randomly chose 51 nuclear genes (see Table 1) for which the orthologous sequences from the human (*Homo sapiens*), mouse (*Mus musculus*), chicken (*Gallus gallus*), and African clawed frog (*Xenopus laevis*) are available in the GenBank. All the protein sequences

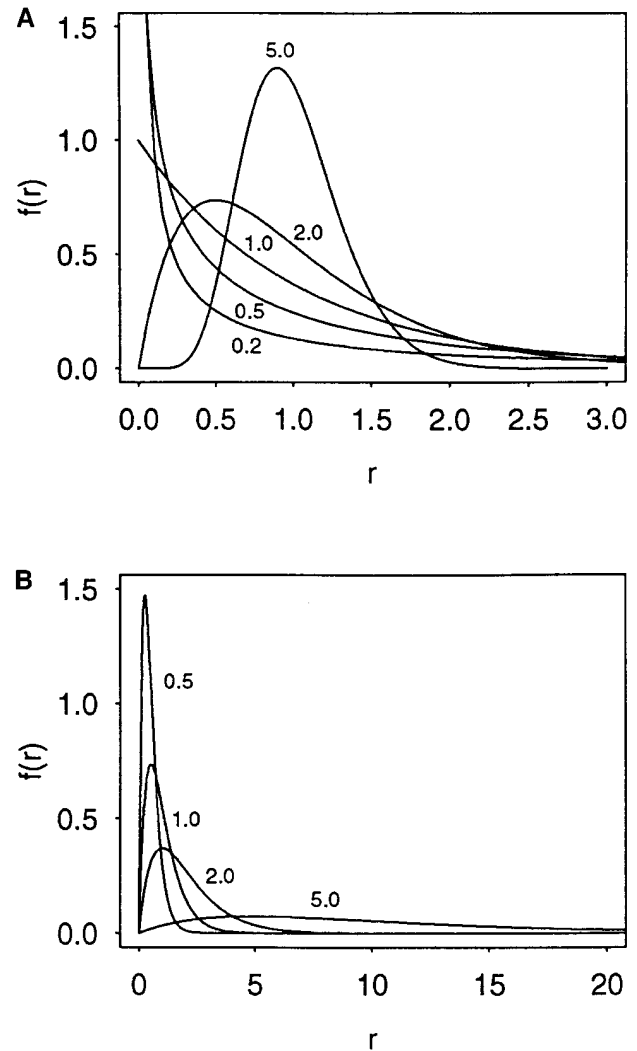


Figure 1.—The probability density functions of gamma distributions. (A) The mean (μ) is fixed at 1 and the shape parameter (α) is presented next to each curve. (B) α is fixed at 2 and μ of each curve is presented.

were downloaded from the HOVERGEN database (Duret *et al.* 1994). Hedges *et al.* (1996) examined the orthology of most of the genes for their purpose of estimating the divergence time of primates and rodents. For the remaining genes, we examined the orthology by phylogenetic analysis as described in Hedges *et al.* (1996). All the sequences used were longer than 100 amino acids. We chose the four species because (1) estimation of the gamma shape parameter α requires at least three sequences, (2) there are many gene sequences available for these species, (3) the phylogeny of the species is known, and (4) the divergence among these species is moderate for most genes so that pairwise distances and substitution rates can be accurately estimated. As shown in Equation 4, to study the mean rate μ , we have to know the total amount of evolutionary time T for a tree. But if we use the same species for all the genes, T will be the same so that we can infer properties of μ from those of μT without knowing the actual value of T .

We also studied 13 mitochondrial genes of 9 mammalian species, including the mouse (GenBank accession no. J01420), rat (*Rattus norvegicus*; X14848), human (D38112), gibbon (*Hylobates lar*; X97707), cow (*Bos taurus*; V00654), whale (*Balaenoptera physalus*; X61145), opossum (*Didelphis virginiana*; Z29573), wallaroo (*Macropus robustus*; Y10524), and platypus (*Ornithorhynchus anatinus*; X83427). In this case only mammalian species were used because mammalian mitochondrial genes evolve quickly (Brown *et al.* 1979) and use of distantly related species may increase estimation errors. The unrooted tree of these 9 species has been well established (*e.g.*, Janke *et al.* 1997).

Sequence analysis: For the orthologous sequences of each gene, we used the known species tree and estimated the number of substitutions (k) along the tree at each amino acid site by using Equation 7 of Gu and Zhang's (1997) method. We then estimated the mean and variance of k . Estimates of μT and $CV(r)$ were then obtained by using Equations 4 and 6, and they were denoted as D and CV , respectively.

We also estimated the gamma shape parameter (α) for the protein sequences of each gene by using Gu and Zhang's (1997) method, which has been shown to be as accurate as, but much faster than, the likelihood method. Actually, the likelihood method (Yang 1994) and the Gu and Zhang's (1997) method gave essentially the same result for the data sets used in this article (data not shown). In Gu and Zhang's (1997) method, the k value for each site is first estimated. Because k follows the negative binomial distribution when r follows the gamma distribution, the gamma shape parameter α is estimated by maximizing the following likelihood function derived from the negative binomial distribution,

$$L(\alpha) = \prod_{i=1}^n \left[\frac{\Gamma(\alpha + k_i) \left(\frac{D}{D + \alpha} \right)^{k_i} \left(\frac{\alpha}{D + \alpha} \right)^\alpha}{k_i! \Gamma(\alpha)} \right], \quad (8)$$

where k_i is the number of substitutions at site i , n is the number of amino acids of the sequence, and $D = \mu T$ is the mean of k_i over all sites, as mentioned earlier. Under the gamma model, the unbiased pairwise distances among the orthologous sequences were estimated by using the formula (Nei *et al.* 1976; Ota and Nei 1994),

$$d = \alpha[(1 - p)^{-1/\alpha} - 1], \quad (9)$$

where d is the gamma distance between two protein sequences, p is the proportional difference, and α is the estimated gamma shape parameter.

For each gene, by a likelihood ratio test, we examined whether the null hypothesis that all sites have the same substitution rate can be rejected and whether the alternative hypothesis of the gamma distribution is favored. We computed

$$l = 2\ln[L(\alpha)/L(\infty)], \quad (10)$$

where α is the estimated gamma shape parameter and $L(\infty)$ is the likelihood of the null hypothesis. Because the null hypothesis is a special case of the alternative hypothesis with α being infinity, l is approximately χ^2 distributed with the number of degrees of freedom being 1.

RESULTS

Nuclear genes: Table 1 shows the estimates for the mean substitution rate and the extent of among-site rate variation for the 51 nuclear genes examined. The coefficient of variation (CV) of the substitution rate among sites varies from 0.43 to 2.02 with the mean of 1.09 and the median of 1.00 (Figure 2). Figure 3 shows the relationships between CV and D (average number of substitutions per site along the whole tree). We fitted the data by a linear regression of CV and $\ln D$. The coefficient of correlation (R) is -0.95 , which is significantly different from 0 ($P < 0.0001$). If we fit the data by a linear regression of CV and D , R becomes -0.91 ($P < 0.0001$). Because $D = \mu T$ and T (total amount of evolutionary time of the tree) is identical for all the genes, the correlation coefficient between CV and D (or $\ln D$) is the same as that between CV and μ (or $\ln \mu$). These results therefore show that the mean substitution rate and the extent of among-site rate variation of a protein are negatively correlated. This observation is independent of any assumption about the underlying distribution of the substitution rate $g(r)$.

We then examined the correlation under the assumption of a gamma distribution for the rate variation among sites. The gamma shape parameter α of the 51 nuclear genes varies from 0.17 to 3.45 with the median of 0.71 (Table 1 and Figure 4). About two-thirds of the genes have α values < 1 , and, therefore, their rate distributions are L-shaped (see Figure 1A). Only 12% of the genes have α values > 2 . Likelihood ratio tests showed that the null hypothesis of no rate variation among sites can be rejected at the 1% significance level for 48 of the 51 genes and can be rejected at the 5% level for the remaining three genes. Therefore, the alternative hypothesis of the gamma distribution is favored. A common feature of the three genes mentioned above is that they are all short (around 150 amino acids), suggesting that the test might not be powerful enough for these genes. The gamma distance d (number of amino acid substitutions per site) between the mouse and chicken varies from 0.057 to 0.703 for the 51 genes (Table 1), which corresponds to the rate of about 0.1 to 1.2 amino acid substitutions per site per billion (10^9) years if we assume that mammals and birds diverged about 300 mya, which is now generally accepted (Benton 1993; Hedges *et al.* 1996).

Figure 5 shows the relationship of α and d (between the mouse and chicken) for the 51 nuclear genes. The R value for α and d is 0.68, which is significantly different from 0 ($P < 0.0001$). Linear regression of α and $\ln d$ does not change the R value very much. Because only

TABLE 1
Substitution rates and among-site rate variations of 51 nuclear genes of vertebrates

Genes	Accession numbers ^a	Amino acids	CV	<i>D</i> ^b	α	<i>p</i> ^c	<i>d</i> ^d
Acetylcholine receptor α subunit	P02708; X03986; X07330; X17244	456	1.28	0.51	0.41	0.158	0.214
Acetylcholine receptor δ subunit	X55019; K02582; K02903; X07069	508	0.83	0.82	1.05	0.315	0.455
Acetylcholine receptor γ subunit	X01715; X03818; K02904; X07068	502	0.70	0.81	1.67	0.323	0.439
α -Globin	V00493; V00714; J00855; X14259	142	0.63	1.02	1.93	0.319	0.425
Annexin II	M14043; M14044; X53334; M60768	339	1.21	0.37	0.55	0.103	0.120
B-myb	X13293; X70472; X67505; M75870	624	0.85	0.94	0.93	0.268	0.370
Bcl2-Ig fusion gene	X06487; M16506; Z11961; X82461	161	0.50	1.54	2.59	0.236	0.284
Beta 2 thyroid hormone receptor	X04707; U15542; M65207; M35359	368	1.98	0.12	0.21	0.049	0.057
C-kit proto-oncogene	X06182; Y00864; D13225; M80798	909	0.44	1.77	3.45	0.343	0.447
C-myb	X52125; M12848; M14129; L22741	615	1.01	0.72	0.68	0.184	0.237
C-yes protein	M15990; X67677; X13207; X14377	535	1.96	0.20	0.20	0.092	0.124
CD18-integrin β -2	X64072; X14951; X71768; S49364	790	1.30	0.42	0.43	0.153	0.203
CDC2	X05360; M38724; X16881; M60680	295	1.50	0.26	0.31	0.092	0.112
Complement component 3	K02765; K02782; U16848; U19253	1493	0.47	1.20	3.29	0.444	0.643
Connexin 43	X52947; X62836; M29003; X17243	378	1.69	0.27	0.24	0.069	0.083
Cyclin A	X68303; Z26580; X72892; X53745	365	0.75	1.28	1.14	0.307	0.432
E-cadherin	Z13009; X06115; M22180; U04708	856	0.70	1.09	1.31	0.379	0.573
Erythroblast virus oncogene homolog 2	J04102; J04103; X07202; X52635	464	1.00	0.73	0.70	0.226	0.310
Estragen receptor	X99101; M38651; X03805; A40907	580	0.97	0.60	0.81	0.203	0.263
Focal adhesion	L13616; M95408; M86656; U11078	1051	2.00	0.20	0.17	0.087	0.119
Follistatin	P19883; Z29532; X87609; P31515	315	1.51	0.33	0.33	0.108	0.136
GATA-3-transcription factor	X55122; X55123; X56931; M76565	430	1.64	0.39	0.24	0.072	0.088
Hedgehog	L38518; X76290; L28099; L35248	408	1.29	0.68	0.36	0.164	0.232
Histon H2A.X	X14850; X58069; V00413; M21287	139	2.02	0.20	0.19	0.085	0.114
Lactate dehydrogenase-A	X02152; U13687; X53828; U07179	330	1.03	0.65	0.71	0.146	0.176
Lamin B1	L37737; X16705; X16878; X06344	577	1.00	0.65	0.82	0.199	0.255
Midkine protein	X55110; M19662; X76482; U06048	140	0.68	0.92	1.63	0.371	0.537
MOS proto-oncogene	J00119; J00620; M19412; X13311	328	0.64	1.12	1.69	0.384	0.561
Muscle pyruvate kinase	X56494; D38379; J00903; U03878	527	1.42	0.28	0.38	0.140	0.186
Myogenic factor Myf-5	X14894; X56182; X73250; X56738	255	0.90	0.88	0.79	0.314	0.482
N-cadherin	X54315; M31131; X07277; X57675	899	1.37	0.46	0.38	0.137	0.179
N-myc	Y00664; X03919; D90071; X58670	404	0.94	0.82	0.77	0.312	0.482
Na ⁺ /K ⁺ ATPase β -1-subunit	M25160; X16646; J02787; U17061	301	0.90	0.78	0.86	0.286	0.412
Na ⁺ /K ⁺ ATPase β -2-subunit	M81181; X16645; L13208; M37788	273	0.62	1.11	2.08	0.487	0.787
NAD(+) ADP-ribosyltransferase	P09874; P11103; P26446; P31669	991	1.00	0.55	0.77	0.218	0.290
Nerve growth factor	X52599; M14805; X04003; X55716	230	0.80	0.90	0.99	0.357	0.555
Neural cell adhesion molecule	X16841; Y00051; M15860; M25696	713	1.10	0.65	0.59	0.138	0.168
Neurotrophin-3	M37763; M53257; M83378; Z30090	228	0.43	1.51	2.45	0.110	0.119
Nucleolar protein N038	M28699; M33212; X17200; X56039	282	0.79	0.86	1.26	0.301	0.415
Nucleolin	M60858; X07699; X17199; X72957	594	0.60	0.95	2.23	0.370	0.514
Ornithine decarboxylase	P11926; P00680; X64710; X56316	448	1.22	0.43	0.47	0.174	0.236
P53 cellular tumor antigen	X02469; X00741; X13057; M36962	343	0.53	1.24	1.81	0.436	0.672
Prohormone convertase	X17094; X54056; Z68093; M80471	590	1.34	0.45	0.40	0.137	0.179
Rhodopsin	U49742; M55171; D00702; S79840	348	1.28	0.32	0.44	0.135	0.172
S6 II kinase	L07597; M28489; M28488; M20187	718	1.80	0.32	0.22	0.079	0.101
SPARC	J03040; X04017; L24906; X62483	296	1.27	0.50	0.42	0.151	0.199
TGF β -2	M19154; X57413; M31154; X51817	409	1.70	0.31	0.22	0.117	0.168
TGF β -3	X14149; M32745; X58071; X51817	389	0.73	0.99	1.04	0.123	0.140
Transcription factor Eryf1	X17254; X15763; M26209; M76566	268	0.62	1.23	1.95	0.452	0.703
Tyrosine phosphatase	X53364; M36033; L22437; U09135	791	1.50	0.51	0.27	0.173	0.275
Vimentin	Z19554; X51438; M15850; X16843	451	1.35	0.43	0.38	0.111	0.138

^a The Gen Bank/EMBL/SwissProt accession numbers are presented in the order of human, mouse, chicken, and frog sequences.

^b Average number of substitutions per site along the tree.

^c Proportional differences between the mouse and chicken proteins.

^d Gamma distances between the mouse and chicken proteins.

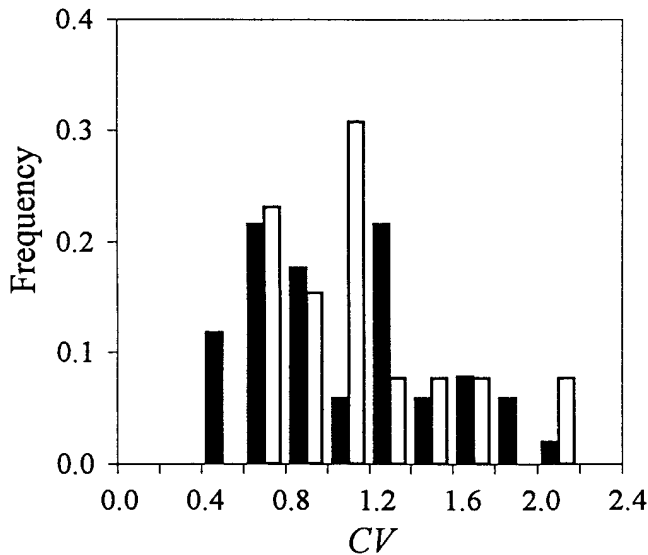


Figure 2.—Distributions of the CV values of 51 nuclear (solid histograms) and 13 mitochondrial (open histograms) genes.

orthologous sequences were used, d is proportional to the average substitution rate μ ($d = 2\mu t$, where t is the time since the divergence of birds and mammals). Therefore, the correlation coefficient between α and μ is also 0.68. Although only the distances between the mouse and chicken are presented in Figure 5, use of pairwise distances among other species gave similar results. Large estimates of α (e.g., $\alpha > 2$) are known to have substantial sampling errors (Gu and Zhang 1997), and they should have lower weights than other data points in the regression. A simple practice is to give

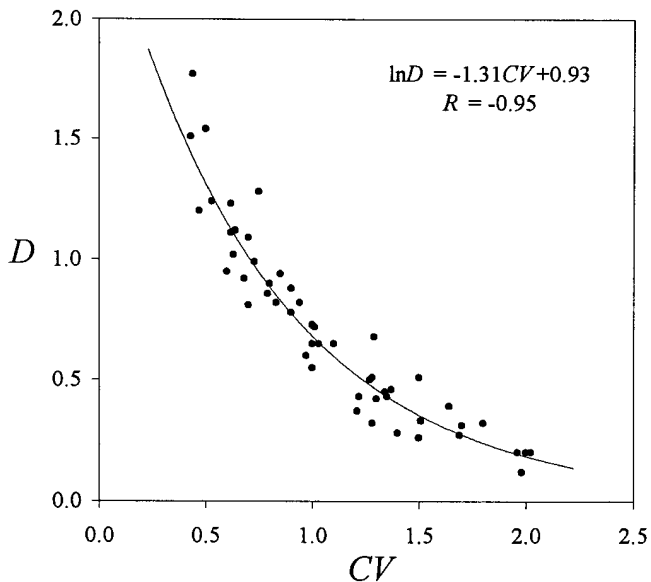


Figure 3.—Correlation between the extent of among-site rate variation CV and the mean substitution rate $D (= \mu T)$ in 51 nuclear genes. R is the coefficient of correlation.

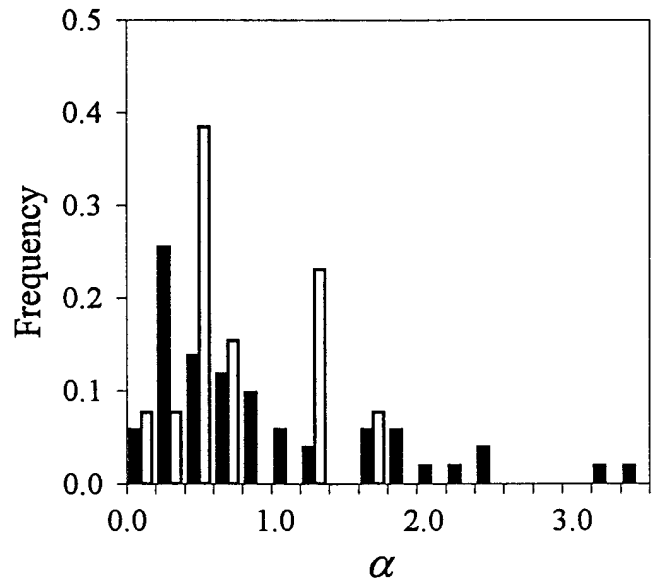


Figure 4.—Distributions of the α values of 51 nuclear (solid histograms) and 13 mitochondrial (open histograms) genes.

zero weights to them. The correlation coefficient for the remaining 45 genes with $\alpha < 2$ is now 0.86 ($P < 0.0001$).

Influence of estimation biases: The above analyses showed that the estimated mean substitution rate and the extent of rate variation among sites are highly correlated. However, to prove that our observations reflect the true properties of protein evolution, we have to

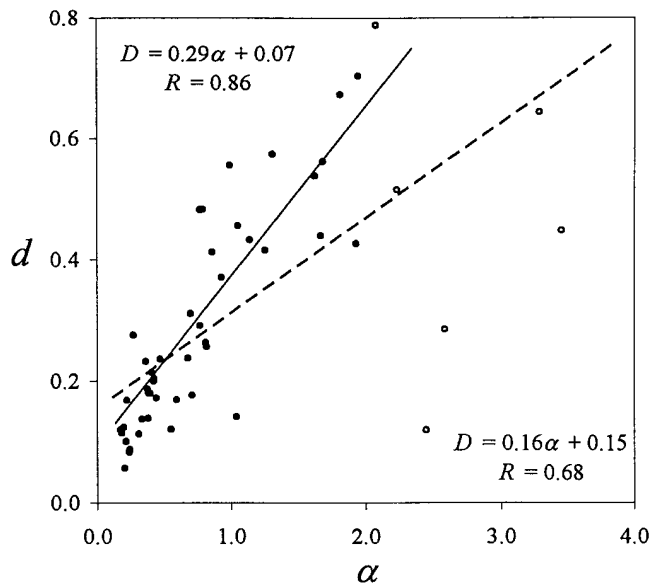


Figure 5.—Correlation between the gamma shape parameter α and the distance d between mouse and chicken proteins. The dashed line is the regression for 51 genes (both solid and open circles), whereas the solid line is the regression for the 45 genes (solid circles) excluding the 6 genes (open circles) whose α values are estimated to be greater than 2. The regression equations presented at the bottom of the figure are for the 51 genes, and the equations at the top are for the 45 genes.

show that the correlation is not an artifact due to some biases of the estimation methods we have used. In the estimation of CV and D , no assumptions of the distribution $g(r)$ were made. It is, therefore, very difficult to examine the estimation biases quantitatively. In the following paragraphs, we studied possible biases in the estimation of α and d , where $g(r)$ was assumed to be gamma.

A computer simulation study (J. Zhang, unpublished results) demonstrated that when the number of sequences used is small (*e.g.*, four), estimation of α is subject to asymptotic bias: α is likely to be underestimated for closely related sequences, but overestimated for highly diverged sequences. This is because when the divergence level is low, there is not enough time for substitutions to occur at potentially variable sites, so they are regarded as invariant sites and α is underestimated. When the divergence level is very high, multiple hits cannot be fully corrected. As a result, the numbers of substitutions are underestimated at very fast evolving sites and α is subsequently overestimated. The estimation bias decreases when the number of sequences used increases (Gu and Zhang 1997; J. Zhang, unpublished results). Because the estimation bias of α is likely to create a positive correlation between α and d (or μ), it is worthwhile to examine the influence of the estimation bias on the present study quantitatively by using computer simulation.

The tree of the human, mouse, chicken, and frog (Figure 6) was used as a model tree in the simulation. The branch lengths of the tree were proportional to the divergence times of the species, that is, 100 million yr between the human and mouse (Hedges *et al.* 1996), 300 million yr between mammals and birds, and 350 million yr between amniotes and frogs (Ahlberg and Milner 1994). In the computer simulation, we simulated the evolution of 45 independent genes. For each gene, we randomly chose an α value from a uniform distribution in the range of 0.2 to 2 as the gamma shape parameter of the gene, and a p value from a uniform

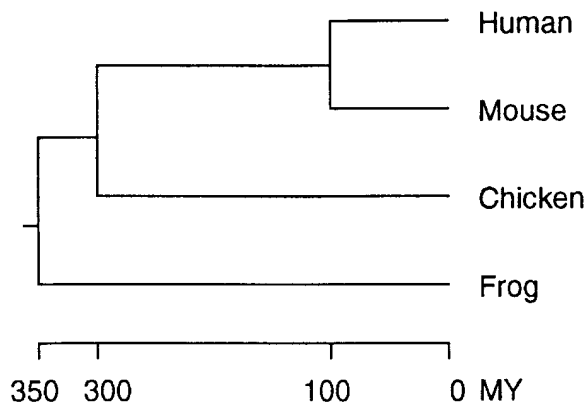


Figure 6.—The model tree used in the simulation for nuclear genes.

distribution in the range of 0.05 to 0.45 as the proportional difference between the mouse and chicken proteins. These ranges represented the observed variations in the 45 genes, excluding the 6 genes with estimated α values >2 . The sequence length used was 450 amino acids, because it is approximately the average length of the proteins that we have examined. The simulation procedure was similar to that described in Zhang and Nei (1997), except that the substitution model used was gamma + Poisson. That is, for each site, the relative substitution rate (r_R) was randomly generated from a gamma distribution with the expectation of 1 and the shape parameter α , and the amino acid substitution followed the Poisson process. The tree branch lengths (expected numbers of substitutions per site) were computed for each site separately. For example, the expected distance between the mouse and chicken is d_{M-C} , computed from Equation 9 by using the α and p values that were randomly chosen for this gene, and the average substitution rate for the protein is then $\mu = d_{M-C} / (2 \times 300)$ substitutions per site per mya, where 300 stands for 300 million yr. The substitution rate at a given site is therefore equal to $r = r_R \mu$, where r_R is the relative rate at the site. The length of the branch from human to the common ancestor of human and mouse is then $100r$ substitutions for the site, where 100 stands for 100 million yr. Other branch lengths can be computed similarly. In the simulation, after the four protein sequences at the exterior nodes of the tree were generated, the gamma shape parameter α was estimated by the method of Gu and Zhang (1997), and the gamma distance (d) between the simulated mouse and chicken sequences was computed by using Equation 9 with the estimated α and p . The simulation results with the estimated $\alpha > 2$ were abandoned. When 45 independent genes were simulated and 45 couples of $\alpha (< 2)$ and d were obtained, the correlation coefficient R was computed. We conducted 2500 replications of such computer simulations, and the distribution of the 2500 R values is presented in Figure 7.

It is seen from Figure 7 that the estimated α and d are positively correlated, though the true α and d have no correlation at all. The correlation coefficient of the estimated α and d is significantly different from 0 because all 2500 simulated R values are positive. The average correlation coefficient is 0.65. However, only 0.32% of the 2500 simulated R values are greater than 0.86, which is the observed R value for the 45 nuclear genes examined. In other words, the observed R value of 0.86 is significantly greater than 0 at the 0.32% significance level, even when the estimation bias of α is considered. Therefore, the observed positive correlation between α and d of the nuclear proteins seems to have underlying causes other than the estimation bias.

In the above simulation, α and p were randomly chosen from a uniform distribution, respectively. But in reality, neither of them follow a uniform distribution

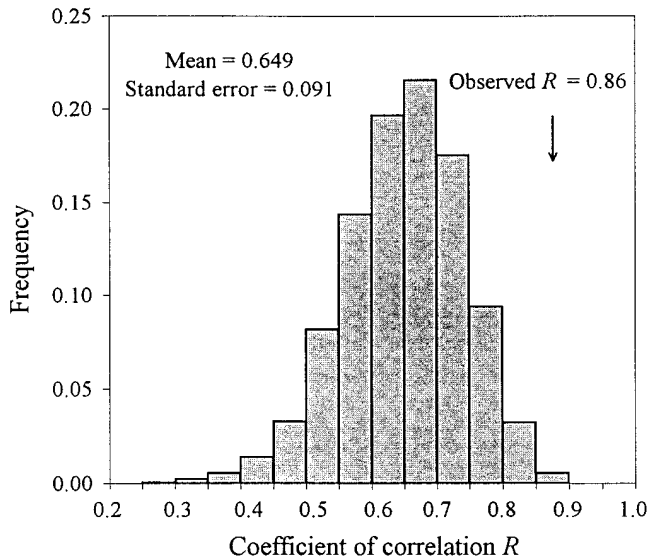


Figure 7.—The distribution of 2500 R values of the estimates of α and μ in the simulation. There is no correlation between the true values of α and μ . The observed R from the 45 nuclear genes is marked.

(see Figure 4). It is expected, and is confirmed in a small-scale simulation, that the R values obtained from the simulation will be smaller when the true distributions of the α and p are used. Therefore, the probability of observing an R value of 0.86 or larger, given that there is no correlation between α and d , may have been overestimated in the simulation. In other words, the effect of the estimation bias may be exaggerated in the simulation and the confidence level of 0.32% is a conservative estimate. The assumptions of the Poisson model of amino acid substitution for protein evolution does not seem to have affected our simulation result very much. Because only orthologous proteins are used in the analysis, the evolutionary rate of a protein is expected to be more or less constant among lineages unless there is variation of the mutation rate due to the generation time effect (Ohta 1995; Li *et al.* 1996), which, however, acts on all the genes to the same extent and does not affect the correlation between CV (or α) and D (or d).

Mitochondrial genes: The estimated CV values of the 13 mitochondrial genes are between 0.66 and 2.05, with the mean of 1.14 and the median of 1.10 (Table 2 and Figure 2). Again, we found that CV and D are highly correlated. The R value is -0.98 ($P < 0.0001$) for the linear regression between CV and $\ln D$ (Figure 8). If we fit the data by a linear regression of CV and D , R becomes -0.92 ($P < 0.0001$). These results show that the mean substitution rate and the extent of among-site rate variation are negatively correlated for the mitochondrial genes as well.

The estimated α values for the mitochondrial genes vary from 0.19 to 1.68 with the median of 0.56 (Table

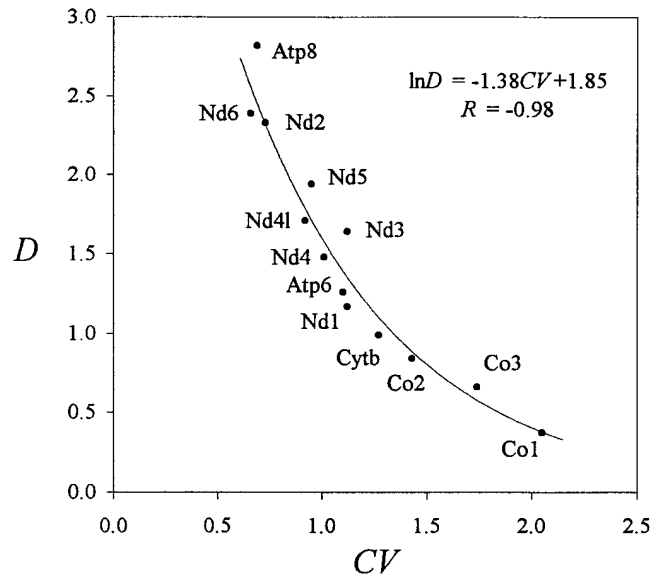


Figure 8.—Correlation between CV and D in 13 mammalian mitochondrial genes.

2; Figure 4). Similar to the nuclear genes, about two-thirds of the mitochondrial genes have α values < 1 . The distance d between the human and mouse varies from 0.20 to 1.15 (Table 2), corresponding to the rate of about 1.0 to 5.8 amino acid substitutions per site per billion years under the assumption that the human and mouse diverged 100 mya. It is seen that mammalian mitochondrial genes evolve about 5 to 10 times faster than nuclear genes, as generally believed (Brown *et al.* 1979).

Figure 9 shows the relationship of α and d (of the

TABLE 2
Substitution rates and among-site variations of 13 mitochondrial genes of mammals

Genes	Amino acids	CV	D^a	α	p^b	d^c
Atp6	226	1.10	1.26	0.51	0.257	0.403
Atp8	60	0.69	2.82	1.38	0.567	1.151
Co1	512	2.05	0.37	0.19	0.092	0.126
Co2	222	1.43	0.84	0.45	0.226	0.445
Co3	261	1.74	0.66	0.24	0.134	0.197
Cytb	379	1.27	0.99	0.44	0.222	0.338
Nd1	315	1.12	1.17	0.56	0.225	0.323
Nd2	345	0.73	2.33	1.31	0.423	0.683
Nd3	113	1.12	1.64	0.42	0.336	0.693
Nd4	457	1.01	1.48	0.64	0.335	0.571
Nd4l	97	0.92	1.71	1.22	0.330	0.474
Nd5	593	0.95	1.94	0.70	0.359	0.621
Nd6	160	0.66	2.39	1.68	0.450	0.718

^a Average number of substitutions per site along the whole tree.

^b Proportional differences between the human and mouse proteins.

^c Gamma distances between the human and mouse proteins.

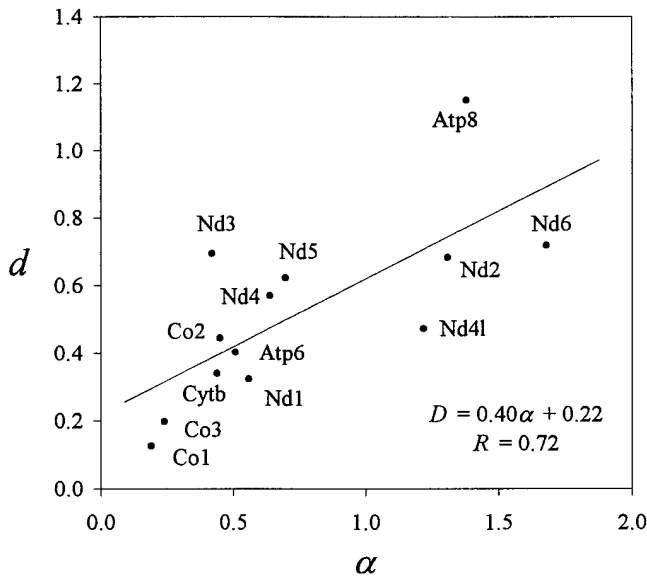


Figure 9.—Correlation between α and the distance d of human and mouse in 13 mammalian mitochondrial genes.

human and mouse) for the mitochondrial genes. The correlation coefficient between α and d (or μ) is 0.72 ($P < 0.01$). Because we have used nine species in the estimation of α , and the estimation bias of α is expected to be quite small when the number of sequences used is no less than eight (Gu and Zhang 1997; J. Zhang, unpublished results), the observed correlation between α and d is unlikely to have been affected by the estimation bias of α . This result was also confirmed by a small-scale computer simulation similar to that performed for the nuclear genes (data not shown). As mentioned before, Kumar (1996) found that the correlation coefficient between α and d for mitochondrial genes is 0.82 by using 1st and 2nd codon positions of the nucleotide sequences. This value is quite close to what we have obtained (0.72), although we used different species and different methods of α estimation. This suggests that the correlation is robust. As in the case of the nuclear genes, the hypothesis of no rate variation among sites was rejected for all 13 mitochondrial genes ($P < 0.01$).

DISCUSSION

Causes of the negative correlation between the mean rate and the extent of among-site rate variation: Through the analysis of 64 genes in total, we have shown that the mean substitution rate and the extent of among-site rate variation of a protein are negatively correlated. We have also excluded the possibility that the correlation is caused by estimation biases or is dependent on the assumption of the gamma distribution of the rate variation among sites. Because the 51 nuclear genes we examined are scattered all over the genome and they encode proteins with a variety of functions, it is unlikely that the correlation is limited to particular chromosomal regions

or is restricted to particular groups of proteins. Rather, the correlation is likely to be a general feature of molecular evolution because both nuclear and mitochondrial genes show similar patterns. Nevertheless, all the data we used are from vertebrates, and our conclusion needs further support from genes of other organisms.

The observation of negative correlation can be explained under the neutral theory of molecular evolution (Kimura 1983; see also Ohta 1992), which asserts that protein sequence evolution is under purifying selection and random genetic drift, and that the rate variation is due to the difference in functional constraints among amino acid sites of a protein and among different proteins. In the following paragraphs, we explain the cause of the correlation by using some simple models of neutral evolution.

Let ν be the mutation rate at an amino acid site, which is assumed to be constant for all sites of all proteins. Then the substitution rate is 0 for an invariant site, between 0 and ν for a variable but functionally constrained site, and equal to ν for a neutral site. Let us first consider a simple model in which the functional constraints at all sites of a protein are equal. Then the mean rate μ and CV have no correlation because CV is always 0, and μ can be any value from 0 to ν . Therefore, this model cannot explain our observation of the correlation between CV and μ . Now let us consider another model, in which there are two types of sites, either invariant or neutral. Then the average substitution rate of the protein becomes

$$\mu = \nu f, \quad (11)$$

where f is the proportion of neutral sites in the protein (Kimura 1983). The variance (Var) and the coefficient of variation (CV) of the rate among sites can be computed as

$$\begin{aligned} \text{Var} &= f(\nu - f\nu)^2 + (1 - f)(0 - f\nu)^2 \\ &= f(1 - f)\nu^2 \end{aligned} \quad (12)$$

and

$$\text{CV} = \frac{\sqrt{\text{Var}}}{\mu} = \frac{\nu\sqrt{f(1-f)}}{f\nu} = \sqrt{\frac{1}{f} - 1}. \quad (13)$$

By using Equations 11 and 13, we obtained

$$\mu = \frac{\nu}{\text{CV}^2 + 1}. \quad (14)$$

Equation 14 shows that, under this model, μ is negatively correlated with CV and is dependent only on CV, when ν is given. Use of more complex models (*e.g.*, Kimura 1979) gives qualitatively similar results. With Equation 14 in mind, we computed the correlation coefficient R for D and $1/(\text{CV}^2 + 1)$. It was 0.95 ($P < 0.0001$) and 0.97 ($P < 0.0001$) for the nuclear and mitochondrial genes, respectively.

In reality, the pattern of the variation of functional

constraint among sites is somewhere between the above two extreme models, and the observation of the strong correlation between μ and CV or $1/(CV^2 + 1)$ suggests that it is probably much closer to the second model than to the first. This is biologically understandable because, for an amino acid sequence to be functional, there must be some essential sites that are invariant and some other sites that are from mildly variable to almost free of any change. This pattern of functional constraints seems to be the requirement for a functional protein. It is unrealistic that the constraints at all amino acid sites of a protein are more or less equal unless the protein is free of any functional constraint (no function). In Equation 14, we can see that ν also influences the relationship of μ and CV. There is evidence that the mutation rate varies among nuclear genes (Wolfe *et al.* 1989). Therefore, the correlation between μ and CV for the nuclear genes may have been depressed due to the variation of mutation rate among genes. This effect is expected to be small for the 13 mitochondrial genes because they have similar mutation rates (Nedbal and Flynn 1998).

The biological meaning of α : The gamma distribution is predominantly used to fit the rate variation among sites of proteins. Our results from the likelihood ratio test indicated that the gamma distribution model fit the rate variation significantly better than the uniform rate model. But the reason why rate variation follows the gamma distribution is unknown and, therefore, the biological meaning of the shape parameter α is not clear. According to our observation of the strong positive correlation between α and μ , α can be interpreted as a measure of functional conservation of a protein. Small values of α indicate strong constraint, and large values indicate weak constraint. It is therefore expected that fourfold degenerate sites and intron regions of functional genes and pseudogenes have very large α values. For example, we estimated that the α of the glycerol-3-phosphate dehydrogenase (GPDH, EC 1.1.1.8) of 13 *Drosophila* and related species (Kwiatowski *et al.* 1997) is about 0.10, but the α of the intron 3 and 5 of the *Gpdh* gene is ~ 3.0 . Actually, the hypothesis of rate constancy over sites cannot be rejected for these intron sequences ($P > 0.05$, goodness of fit test). However, the above interpretation of α is valid only when the majority of amino acid sites are undergoing neutral evolution (*i.e.*, no positive selection). When strong positive selection is acting on a large number of sites, μ tends to be high and α tends to be small because functionally most important sites remain invariant, whereas positively selected sites have very high substitution rates. So, in this case, α does not measure the functional conservation of a protein. For example, we found that the α value of the sperm lysin of 20 abalone species (Lee *et al.* 1995) is 0.16, whereas the μ is about 2 to 3 times the mutation rate ν . Here, ν was estimated under the assumption that the mutation rate per amino acid is 2.25 times the

mutation rate per nucleotide (Nei 1975, p. 225), which was further assumed to be equal to the rate of synonymous nucleotide substitution. So, if this gene is presented in Figure 5, it will sit at the top left corner, far from the regression line.

It is intriguing that there are no dots sitting in the upper left corner in Figure 5 or the upper right corner in Figure 3. These are the locations where both the mean rate and the among-site rate variation are high, and where we expect to find proteins consisting of a number of conserved sites and a number of directionally selected sites. The simplest explanation would be that such genes are not included in our sample of 64 genes. Because our sample is random and is not small, the result suggests that those genes that are under strong positive selection are rare in the genome. Of course, we cannot distinguish neutral genes from those in which positive selection operates in a very small number of sites, because in the latter, the overall relationship between α and μ is mainly determined by the sites that are neutral or under purifying selection. In fact, there are no statistical methods that can effectively detect positive selection that only acts on one or several sites except in the case of convergent and parallel evolution (Zhang and Kumar 1997). It has to be noted that all we have studied here is orthologous gene evolution. Because orthologous genes usually have the same function, it is understandable that we did not find signals of strong positive selection (Sharp 1997). Strong selection is more likely to operate on newly duplicated genes during the evolution of novel gene functions (Zhang *et al.* 1998).

It is also interesting that no dots were found in the lower left corner of Figure 3, where both the mean rate and the rate variation are small. This suggests that even in the most conserved proteins it is unlikely that all sites are equally conserved. More probably, there are still a few nearly neutral sites so that CV becomes high (see Equation 13).

Other implications: Accelerated evolution in some evolutionary lineages has been observed in several proteins (*e.g.*, Stewart *et al.* 1987; Tucker and Lundrigan 1993; Wallis 1993, 1996; Whitfield *et al.* 1993). If the mutation rate is constant among lineages, there are two possible causes of the enhanced substitution rate. One is relaxation of functional constraint, and the other is positive selection. When the rate of nonsynonymous nucleotide substitution is not significantly greater (or even not greater) than the rate of synonymous substitution, it is often difficult to distinguish between the two possible causes of the accelerated evolution (*e.g.*, Li and Gojobori 1983). The relationship between μ and α may provide us some information, because when μ is high, high values of α are expected if functional constraint is reduced, but small values of α are likely if strong positive selection acts on a large number of sites (as in the case of the abalone sperm lysin).

Using the relationship between μ and α , let us examine an interesting case of fast evolution. Whitfield *et al.* (1993) and Tucker and Lundrigan (1993) reported that the sex-determining protein SRY evolves very fast in primates and rodents, and the rates of nonsynonymous and synonymous nucleotide substitution are similar. But because the nonsynonymous rate is not significantly higher than the synonymous rate, these authors were not able to determine whether the accelerated evolution is due to positive selection or relaxation of functional constraints. SRY protein is a transcription factor that binds to specific DNA sites to regulate the expression of its downstream genes in the sex determination pathway. SRY consists of N-terminal, HMG, and C-terminal domains (with 58, 78, and 68 amino acids, respectively, in the human sequence). The HMG domain is responsible for DNA binding and is very conservative, whereas the other two domains are highly variable, particularly the C domain. The length of the HMG domain remains the same among placental mammals, but the lengths of the other two domains vary within or among mammalian orders. The estimated α values of the SRY proteins of eight primate species (Whitfield *et al.* 1993) are 1.58, 0.24, and 3.73 for the N, HMG, and C domains, respectively. The large values of α for the N and C domains, particularly the latter, suggest that the accelerated evolution in SRY is probably due to low functional significance out of the HMG domain rather than adaptation under strong positive selection. Actually, all clinical mutations in human SRY resulting in phenotypic sex reversal are found in the HMG domain (Hawkins *et al.* 1992; Goodfellow and Lovell-Badge 1993; Werner *et al.* 1995), except for one case in which a nonsense mutation happened in the C domain (Tajima *et al.* 1994). Pontiggia *et al.* (1995) found that SRY proteins from different primate species are very similar in both DNA binding and bending and predicted that they should be able to substitute for each other.

Ignoring rate variation among sites causes underestimation of distances between homologous proteins. For estimating dates of early divergences by the molecular clock, only slowly evolving proteins can be used. Because slowly evolving proteins tend to have high degrees of rate variation among sites, consideration of the rate variation and use of gamma distances seems to be necessary for such estimations. This has been highlighted in recent debates on the age of the most recent common ancestor of present-day organisms (Doolittle *et al.* 1996; Hasegawa *et al.* 1996; Miyamoto and Fitch 1996; Feng *et al.* 1997; Gu 1997). In this case, neglect of the rate variation among sites was found to bias the time estimation at least 0.5–1 billion years. However, it is worth mentioning that use of the gamma distance enlarges the sampling variance so that many genes are needed for obtaining reliable estimates of divergence times.

We are grateful to Sudhir Kumar and Patrick Parker for provid-

ing us aligned protein sequences of the nuclear genes. We thank Sudhir Kumar, Masatoshi Nei, Tanya Sitnikova, Chung-I Wu, and two anonymous reviewers for their comments on an earlier version of this article. This work was supported by the research grants from the National Institutes of Health and the National Science Foundation to M. Nei.

LITERATURE CITED

- Ahlberg, P. E., and A. R. Milner, 1994 The origin and early diversification of tetrapods. *Nature* **368**: 507–514.
- Benton, M. J., 1993 *The Fossil Record*, Vol. 2, Chapman and Hall, New York.
- Brown, W. M., M. George, Jr. and A. C. Wilson, 1979 Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **76**: 1967–1971.
- Doolittle, R. F., D.-F. Feng, S. Tsang, G. Cho and E. Little, 1996 Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271**: 470–477.
- Duret, L., D. Mouchiroud and M. Gouy, 1994 HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* **22**: 2360–2365.
- Feng, D.-F., G. Cho and R. F. Doolittle, 1997 Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl. Acad. Sci. USA* **94**: 13028–13033.
- Golding, G. B., 1983 Estimation of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol.* **1**: 125–142.
- Goodfellow, P. N., and R. Lovell-Badge, 1993 SRY and sex determination in mammals. *Annu. Rev. Genet.* **27**: 71–92.
- Gu, X., 1997 The age of the common ancestor of eukaryotes and prokaryotes: statistical inferences. *Mol. Biol. Evol.* **14**: 861–866.
- Gu, X., and J. Zhang, 1997 A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* **14**: 1106–1113.
- Gu, X., and W.-H. Li, 1998 Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *Proc. Natl. Acad. Sci. USA* **95**: 5899–5905.
- 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.
- Hasegawa, M., W. M. Fitch, J. P. Gogarten, L. Olendzenski, E. Hilario *et al.*, 1996 Dating the cenacestor of organisms. *Science* **274**: 1750–1753.
- Hawkins, J. R., A. Taylor, P. Berta, J. Leveilliers, B. van der Auwera *et al.*, 1992 Mutational analysis of SRY: nonsense and missense mutations in XY sex reversal. *Hum. Genet.* **88**: 471–474.
- 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.
- Janke, A., X. Xu and U. Arnason, 1997 The mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl. Acad. Sci. USA* **94**: 1276–1281.
- Kimura, M., 1979 Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl. Acad. Sci. USA* **76**: 3440–3444.
- Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Kimura, M., and T. Ohta, 1974 On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA* **71**: 2848–2852.
- Kumar, S., 1996 Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* **143**: 537–548.
- Kwiatowski, J., M. Krawczyk, M. Jaworski, D. Skarecky and F. J. Ayala, 1997 Erratic evolution of glycerol-3-phosphate dehydrogenase in *Drosophila*, *Chymomyza*, and *Ceratitis*. *J. Mol. Evol.* **44**: 9–22.
- Lee, Y.-H., T. Ota and V. D. Vacquier, 1995 Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* **12**: 231–238.
- Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Li, W.-H., and T. Gojobori, 1983 Rapid evolution of goat and sheep

- globin genes following gene duplication. *Mol. Biol. Evol.* **1**: 94–108.
- Li, W.-H., D. L. Ellsworth, J. Krushkal, B. H.-J. Chang and D. Hewett-Emmett, 1996 Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5**: 182–187.
- 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.
- Nedbal, M. A., and J. J. Flynn, 1998 Do the combined effects of the asymmetric process of replication and DNA damage from oxygen radicals produce a mutation-rate signature in the mitochondrial genome? *Mol. Biol. Evol.* **15**: 219–223.
- Nei, M., 1975 *Molecular Population Genetics and Evolution*. North-Holland Publishing Company, Amsterdam.
- 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.
- Ohta, T., 1992 The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**: 263–286.
- Ohta, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**: 56–63.
- 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.* **36**: 642–643.
- Pontiggia, A., S. Whitfield, P. N. Goodfellow, R. Lovell-Badge and M. E. Bianchi, 1995 Evolutionary conservation in the DNA-binding and -bending properties of HMG-boxes from SRY proteins of primates. *Gene* **154**: 277–280.
- Sharp, P. M., 1997 In search of molecular Darwinism. *Nature* **385**: 111–112.
- Stewart, C.-B., J. W. Schilling and A. C. Wilson, 1987 Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**: 401–404.
- Tajima, T., J. Nakae, N. Shinohara and K. Fujieda, 1994 A novel mutation localized in the 3' non-HMG box region of the SRY gene in 46X, Y gonadal dysgenesis. *Hum. Mol. Genet.* **3**: 1187–1189.
- Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in human and chimpanzees. *Mol. Biol. Evol.* **10**: 512–526.
- Tucker, P. K., and B. L. Lundrigan, 1993 Rapid evolution of the sex determining locus in Old World mice and rats. *Nature* **364**: 715–717.
- Uzzell, T., and K. W. Corbin, 1971 Fitting discrete probability distributions to evolutionary events. *Science* **172**: 1089–1096.
- Wallis, M., 1993 Remarkable high rate of molecular evolution of ruminant placental lactogens. *J. Mol. Evol.* **37**: 86–88.
- Wallis, M., 1996 The molecular evolution of vertebrate growth hormones: a pattern of near-stasis interrupted by sustained bursts of rapid change. *J. Mol. Evol.* **43**: 93–100.
- Werner, M. H., J. R. Huth, A. M. Gronenborn and G. M. Clore, 1995 Molecular basis of human 46X, Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell* **81**: 705–714.
- Whitfield, L. S., R. Lovell-Badge and P. N. Goodfellow, 1993 Rapid sequence evolution of the mammalian sex-determining gene SRY. *Nature* **364**: 713–715.
- Wolfe, K. H., P. M. Sharp and W.-H. Li, 1989 Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283–285.
- Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**: 306–314.
- Zhang, J., and S. Kumar, 1997 Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* **14**: 527–536.
- Zhang, J., and M. Nei, 1997 Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* **40** (Suppl. 1): S139–S146.
- Zhang, J., H. F. Rosenberg and M. Nei, 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* **95**: 3708–3713.

Communicating editor: C.-I Wu