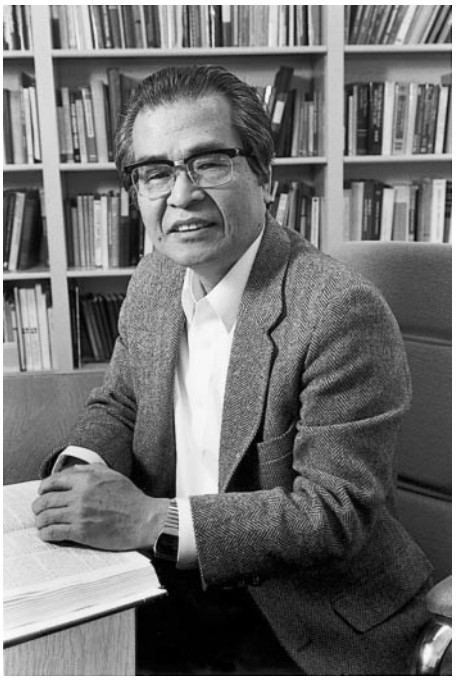
The Wilhelmine E. Key 2001 Invitational Lecture

Estimation of Divergence Times for a Few Mammalian and Several Primate Species

M. Nei and G. V. Glazko



Masatoshi Nei is Evan Pugh Professor of Biology and Director of the Institute of Molecular Evolutionary Genetics at Pennsylvania State University, University Park. He received his Ph.D. in quantitative genetics from Kyoto University, Japan, and then did postdoctoral work at the University of California at Davis and North Carolina State University. He was president of the Society of Molecular Biology and Evolution and the American Genetic Association and served on the editorial board of many scientific journals. With Walter Fitch, he created the journal *Molecular Biology and Evolution* and started the Society for Molecular Biology and Evolution. He authored or edited six books and has written more than 230 scientific papers. He is one of the original members of "Highly Cited Researchers" at the Institute for Scientific Information, Philadelphia. He is a member of the National Academy of Sciences and is a fellow of the American Academy of Arts and Sciences and the American Association of Advancement of Science. This (Wilhelmine E. Key) lecture was delivered on May 19, 2001, at the American Genetic Association Symposium, "Primate Evolutionary Genetics," at the Town and Country Resort and Convention Center, 500 Hotel Circle South, San Diego, CA 92108. This work was supported by research grants from the National Institutes of Health (GM20293) and the National Aeronautic and Space Administration (NCC2-1057) to M.N. From the Department of Biology, 328 Mueller Laboratory, Pennsylvania State University, University Park, PA 16802. Address correspondence to Masatoshi Nei at the address above or e-mail: nxm2@psu.edu.

 93:157–164

Statistical methods for estimating divergence times by using multiprotein gamma distances are discussed. When a large number of proteins are used, even a small degree of deviation from the molecular clock hypothesis can be detected. In this case, one may use the stem-lineage method for estimating divergence times. However, the estimates obtained by this method are often similar to those obtained by the linearized tree method. Application of these methods to a dataset of 104 proteins from several vertebrate species indicated that the divergence times between humans and mice and between mice and rats are about 96 and 33 million years (MY) ago, respectively. These estimates were obtained by assuming that birds and mammals diverged 310 MY ago. Similarly application of the methods to the protein sequence data from primate species indicated that the human lineage separated from the chimpanzee, gorilla, Old World monkeys, and New World monkeys about 6.0, 7.0, 23.0, and 33.0 MY ago, respectively. In this case the use of two calibration points, that is, the divergence time (13 MY ago) between humans and orangutans and between primates and artiodactyls (90 MY ago) gave essentially the same estimates.

According to the molecular clock hypothesis, the number of amino acid substitutions in a protein increases roughly in proportion to the time since divergence of the two species compared (Margoliash 1963; Zuckerkandl and Pauling 1962). Strictly speaking, no gene or protein would evolve at a constant rate for a long evolutionary time, because gene function is likely to change over time (Nei and Kumar 2000) and the mutational and DNA repair mechanisms appear to vary among different groups of organisms (Britten 1986). However, even if the substitution rate is not strictly constant, it is still possible to obtain rough estimates of divergence times, and these estimates are very useful when there is no reliable fossil record (Wilson *et al.* 1977). Furthermore, if a gene evolves excessively fast or slow in a few evolutionary lineages, one can eliminate these lineages and estimate divergence times for the rest of the species (Takezaki *et al.* 1995). The accuracy of time estimates is expected to increase as the number of genes or proteins used increases, and in recent years many authors have used multiple genes or proteins for this purpose (e.g., Doolittle *et al.* 1996; Kumar and Hedges 1998; Murphy *et al.* 2001; Wray *et al.* 1996).

There are several statistical methods for estimating divergence times, but the theoretical basis of the methods is not well understood when multiple genes are used. Nei *et al.* (2001) recently examined the reliability of different methods for estimating divergence times and reached the conclusion that the phylogenetic tree for the species to be used first should be established and the divergence times estimated by using the multiprotein (or multigene) or the weighted average gamma distances (see below). These methods are called the concatenated distance (CD) approach. This is different from the commonly used individual protein or gene (IP or IG) approach, in which the divergence time for a pair of species is estimated from each protein and the average of the times for all proteins is used as the final estimate.

In this article we present the computational procedures of the CD and IP methods in some detail and apply the methods for estimating the divergence times for a few mammalian and several primate species.

Methods of Estimating Divergence Times

Individual Protein (IP) Approach

In the past, most investigators have used a method that is called the IP approach

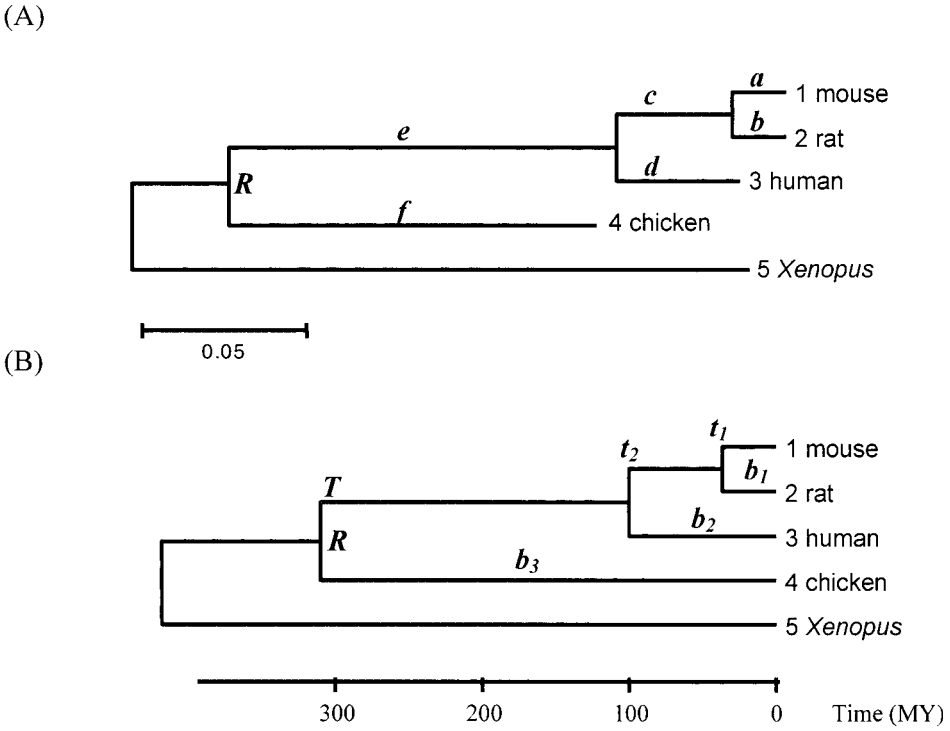


Figure 1. Phylogeny of the five vertebrate species used. **(A)** NJ tree constructed using multiprotein gamma distance (d_{MG}) for 104 protein sequences. **(B)** Linearized tree. *R*: root.

(Ayala *et al.* 1998; Feng *et al.* 1997; Gu 1998; Kumar and Hedges 1998; O'hUigin and Li 1992). In this approach, the estimate of divergence time is computed for each protein (or gene) and the average of the estimates over all proteins is used as the final estimate. Consider Figure 1A, in which a phylogenetic tree for five species is given. Here, *a, b, c, d, e,* and *f* stand for the least-squares estimates of branch lengths (number of amino acid substations) for a protein. Species 5 is used as an outgroup to determine the root of the tree for the remaining sequences, and therefore the branch length estimate for this branch is not given. Here we assume that the topology of the tree for the five species has been established from other information. To estimate divergence times between species, it is convenient to construct a linearized tree (Takezaki *et al.* 1995), in which the branch lengths are reestimated under the assumption of a constant-rate evolution (Figure 1B). When this linearized tree is constructed, a timescale for the tree is produced to estimate divergence times (t_1 and t_2). This timescale can be obtained by computing the rate of amino acid substitution per year (r) by using the known divergence time and the corresponding branch length estimate for a pair of species or species clusters.

For example, if T is the calibration point in Figure 1B, the rate of amino acid substitution can be estimated by $\hat{r} = \hat{b}_3/T$, where $\hat{b}_3$ is the branch length estimate for species 4 after divergence from species 1, 2, and 3 in the linearized tree (Figure 1B). (When there are two or more calibration points, r can also be obtained by the regression coefficient method; Takahashi *et al.* 2000). When this rate is obtained, we can estimate t_1 by

$$\hat{t}_1 = \hat{b}_1/\hat{r} = (\hat{b}_1/\hat{b}_3)T. \qquad (1)$$

Here the estimates $\hat{b}_1$, $\hat{b}_2$, and $\hat{b}_3$ can be obtained from pairwise distances by using Takezaki *et al.*'s (1995) method. Similarly the estimate of t_2 is given by $\hat{t}_2 = \hat{b}_2/\hat{r} = (\hat{b}_2/\hat{b}_3)T$. The variances of $\hat{t}_1$ or $\hat{t}_2$ can be obtained by Takezaki *et al.*'s method or by the bootstrap method (Su and Nei 1999). Before the construction of a linearized tree, it is customary to conduct a statistical test of the molecular clock hypothesis and eliminate species that evolved excessively fast or slow. This test can be done by using Takezaki *et al.*'s U statistic.

When there are data from n different proteins, the average of $\hat{t}_1$ s for all proteins is used as an estimate of $\hat{t}_1$. That is,

$$\hat{t}_1 = \left[\frac{\hat{b}_{11}}{\hat{b}_{31}} + \frac{\hat{b}_{12}}{\hat{b}_{32}} + \cdots + \frac{\hat{b}_{1k}}{\hat{b}_{3k}}\right]\frac{T}{k}, \qquad (2)$$

where $\hat{b}_{1i}$ and $\hat{b}_{3i}$ are estimates of b_1 and b_3, respectively, for the ith protein, and k is the total number of proteins used. Theoretically, however, $\hat{t}_1$ obtained in this way tends to give biased estimates of t_1, even if the branch length estimates (numbers of amino acid substitutions) for each protein are unbiased (Nei *et al.* 2001). Particularly when one knows t_1 and wants to estimate T, where $T > t_1$, the bias of the estimate tends to be upward.

Although the Poisson correction (PC) distance (Nei and Kumar 2000) appears to give sufficiently accurate estimates of divergence times (Nei 1987) when protein sequences are closely related, the time estimates are usually obtained by using the following PC gamma distance

$$d = a[(1 - p)^{-1/a} - 1], \qquad (3)$$

where a is the shape parameter of the gamma distribution (gamma parameter) and decreases as the variation of r among sites increases (Nei *et al.* 1976; Ota and Nei 1994). A convenient way of estimating a is to use Gu and Zhang's (1997) method.

If $\hat{a}$ is an estimate of a rather than a constant, the variance of d is given by

$$V(d) = V_1(d) + V(\hat{a})\left[(1 - p)^{-1/a} \times \left\{1 + \frac{1}{a}\ln(1 - p)\right\} - 1\right]^2, \qquad (4)$$

where $V_1(d) = p(1 - p)^{-(1+2/a)}$, $V(\hat{a}) = [2a(a + 1)(p + a)^2]/(np^2)$, and n is the number of amino acids used (Nei *et al.* 2001).

One might question the applicability of Equation 3 to actual data, because it does not take into account higher rates of substitution between similar amino acids than between dissimilar amino acids (Dayhoff *et al.* 1978). Grishin (1995) developed a complex distance measure by taking into account the variation in substitution rate among different amino acid sites as well as among different pairs of amino acids. However, this distance can be approximated very well by a PC gamma distance with $a = 0.65$ (Nei and Kumar 2000). Dayhoff's PAM distance (Dayhoff 1978) can also be computed by Equation 3 with $a = 2$ (Ota and Nei 1994) or $a = 2.25$ (Nei and Kumar 2000). Therefore, for most practical purposes, we may use PC gamma distance.

Theoretically one might think that the a value should be estimated from closely related sequences, because these sequences would provide the estimate based on the

instantaneous rate of amino acid substitution. In practice, however, the estimate obtained in this way appears to be generally unduly low to be used for estimating the divergence times of distantly related sequences (Zhang and Gu 1998; Nei *et al.* 2001). Apparently the evolutionary pattern of distantly related sequences should be described differently from that of closely related sequences.

Concatenate Distance (CD) Approach

Previously we mentioned that to obtain an unbiased estimate of t_1, pairwise concatenate distances for all proteins should be computed and b_1 and b_3 estimated from these distances. There are several ways of concatenating pairwise distances for different proteins to obtain unbiased estimates of b_1 and b_3, but Nei *et al.*'s (2001) study has suggested that (1) multiprotein (or multigene) gamma distance (d_{MG}) and (2) weighted average distance (d_{WA}) are appropriate for this purpose. The d_{MG} is a PC or PC gamma distance for the concatenated amino acid sequence for all proteins, whereas d_{WA} is an average of gamma distance for different proteins weighted with sequence length. The variances of d_{MG} and d_{WA} are obtained by the bootstrap or the jackknife method by using genes as the units of resampling.

Divergence Times Between Mice and Rats and Between Humans and Rodents

A large number of authors have estimated the times of divergence between different groups of mammals by using molecular data (Dickerson 1971; Eizirik *et al.* 2001; Kumar and Hedges 1998; Li *et al.* 1990), but the results obtained are conflicting and controversial (Arnason *et al.* 1996; Bromham *et al.* 1999; Foote *et al.* 1999). Of special interest in this regard are the divergence times between mice and rats and between humans and rodents. Molecular estimates of these divergence times have been controversial because the fossil record is poor (Benton 1993; Easteal 1999) and rodent genes appear to have evolved faster than primate genes (Gu and Li 1992). In this article we therefore focus our attention first on these divergence times. We use five vertebrate species—mice, rats, humans, chickens, and *Xenopus laevis*—in which the evolutionary relationships are well established and for which many shared protein sequences are available. *Xenopus* is used as an outgroup species (Figure 1).

We obtained protein sequence data from the HOVERGEN database (Duret *et al.* 1994). Using the procedure described by Nei *et al.* (2001), we then secured 104 putative orthologous proteins with a total of 48,092 amino acids (see http://www.bio.psu.edu/people/faculty/nei/lab) and computed d_{MG} and d_{WA} for all pairs of species. The gamma parameter (a) for these distances was estimated by Gu and Zhang's (1997) method (the computer program is available on the website at http://mep.bio.psu.edu).

The neighbor-joining (NJ) tree (Saitou and Nei 1987) and the branch length estimates obtained by using d_{MG} are presented in Figure 1A. Application of Takezaki *et al.*'s (1995) U statistic for the group of mouse, rat, human, and chicken has shown that the molecular clock does not hold and that chicken genes evolved significantly slower than mammalian genes and rodent genes evolved significantly faster than human genes. Note that in the present case even a small extent of rate variation can be detected because a large number of amino acids are used. However, because a certain degree of deviation from the molecular clock hypothesis does not disturb time estimates seriously (Nei and Kumar 2000), we constructed a linearized tree (Figure 1B) to obtain rough estimates of divergence times for mammalian species. We did this under the assumption that the chicken and mammalian lineages diverged 310 million years (MY) ago (Benton 1993; Kumar and Hedges 1998). The results suggest that the human and rodent lineages diverged about 96 MY ago and mice and rats diverged about 33 MY ago (d_{MG} in Table 1). Interestingly, these estimates are close to those obtained by a different method mentioned below.

Theoretically it would be better to consider the variation of evolutionary rate among lineages in the estimation of divergence times. The problem is that we usually do not know how the evolutionary rate changed with time. In the absence of this knowledge, the divergence times t_1 and t_2 may be estimated by assuming that the branch length between the branch point (R) of the chicken and the mammalian lineages and the average exterior node of rodents, which is given by $L = [(a + b)/2 + c + e)]$ in Figure 1A, corresponds to 310 MY. In this case t_1 and t_2 are estimated by $\hat{t}_1 = [(a + b)/2L)]T$ and $\hat{t}_2 = [(a + b + 2c)/(2L)]T$, respectively. We call this approach the stem-lineage method, and it gives $\hat{t}_1 = 31$ MY and $\hat{t}_2 = 96$ MY (d_{MG} in Table 1). These values are very close to those obtained by the linearized-tree method, suggesting that the linearized-tree method gives quite reasonable results in the present case. Note that the stem-lineage method can be used only when the total number of amino acids is so large that the standard errors of branch length estimates are small, as in the present case.

In Table 1, the time estimates obtained by using the weighted average distance (d_{MA}) are also presented. These estimates are quite close to those obtained using d_{MG}. Therefore we can use any of d_{MG} and d_{WA} for estimating divergence times, but it is easier to compute d_{MG} than d_{WA}. The variance of d_{MG} in Table 1 was computed by the jackknife method using genes as the units of resampling, whereas the variance of d_{WA} was computed by the bootstrap because of simplicity.

Previously we mentioned that the traditional IP method tends to give biased estimates of divergence times. The estimates of t_1 and t_2 obtained by this method are given in Table 1. They are certainly considerably higher than those obtained by the CD methods. In the present case, the slow evolutionary rate of the chicken sequence also contributed to the higher values of $\hat{t}_1$ and

Table 1. Estimates (± standard errors) of divergence times (MY) between mice and rats and between humans and rodents

	Mouse–rat (t)		Human–rodent (t_2)	
Method	Linearized[a]	Stem-lineage[b]	Linearized	Stem-lineage
Concatenate distance (CD) approach				
d_{MG} (a = 0.57)	32.9 ± 2.3	29.4 ± 2.1	95.5 ± 4.2	94.0 ± 4.4
d_{WA} ($\bar{a}$ = 0.76)	33.0 ± 2.0	30.6 ± 2.9	97.6 ± 4.4	96.3 ± 5.0
d_D (a = 2.25)	38.3 ± 2.7	35.0 ± 2.5	106.3 ± 4.6	104.4 ± 4.6
Individual protein (IP) approach				
d_G ($\bar{a}$ = 0.76)	38.5 ± 3.2		102.9 ± 5.0	
d_D (a = 2.25)	43.5 ± 3.4		107.9 ± 4.7	

One hundred four genes with a total of 48,092 amino acids were used.

[a] Linearized tree method.

[b] Stem-lineage estimation method.

d_{MG}, multiprotein distance; d_{WA}, weighted average distance; d_D, Dayhoff PAM distance.

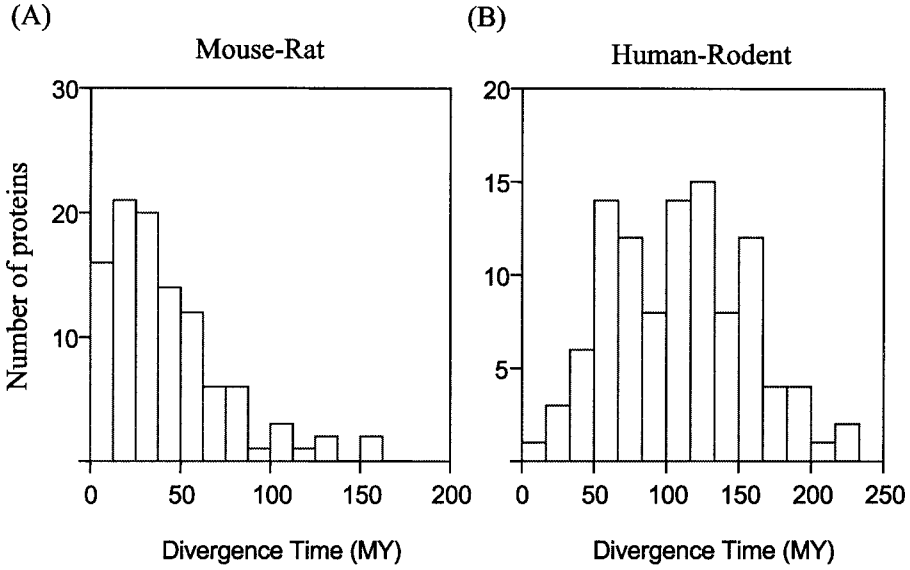


Figure 2. Distribution of single-protein estimates of the divergence time between (**A**) mice and rats and (**B**) humans and rodents. The total number of proteins used is 104.

$\hat{t}_2$ obtained by the IP approach. These explain the higher divergence times ($\hat{t}_1$ = 41 MY and $\hat{t}_2$ = 112 MY) obtained by Kumar and Hedges (1998), who used the IP approach.

Figure 2 shows the histograms of $\hat{t}_1$ and $\hat{t}_2$ obtained from individual proteins. The distribution of $\hat{t}_1$ and $\hat{t}_2$ are quite wide, indicating that the estimates of t_1 and t_2 based on a small number of proteins are unreliable. Ideally we should use as many genes as possible, but at least about 20 genes.

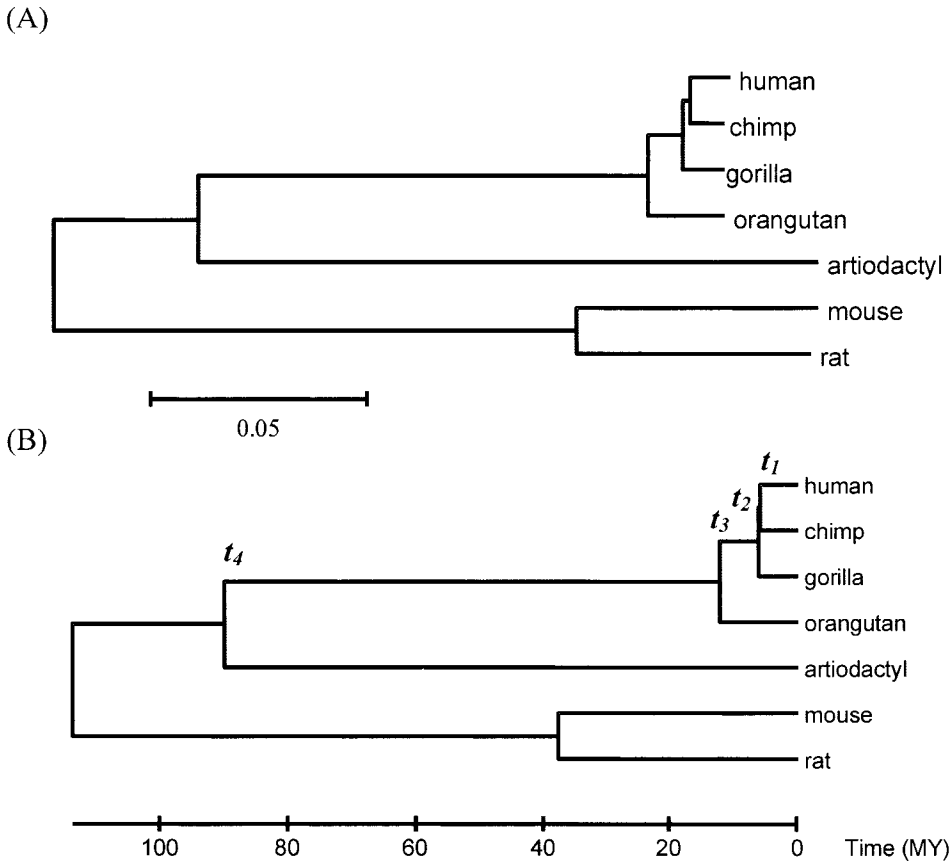


Figure 3. **(A)** NJ tree for hominoid species constructed using multiprotein gamma distance (d_{MG}) with a = 0.54 for 27 protein sequences. **(B)** Linearized tree.

Divergence Times for Several Species of Higher Primates

Another controversial subject with respect to time estimates is that of divergence times for higher primates. The tree topology for major groups of higher primates, that is, hominoids, Old World (OW) monkeys, and New World (NW) monkeys, is fairly well established (Goodman *et al.* 1998). However, the divergence times of these species remain controversial (e.g., Arnason *et al.* 1998; Cao *et al.* 2000; Horai *et al.* 1995; Takahata and Satta 1997). Many of these studies have been conducted using mitochondrial DNA (mtDNA), but mtDNA evolves so erratically that the estimates obtained using mtDNA appear unreliable (Gissi *et al.* 2000; Glazko GV and Nei M, unpublished data). We have therefore examined the divergence times of these species by using nuclear proteins.

Four Hominoid Species

We first considered four hominoid species (humans, chimpanzees, gorillas, and orangutans) and cattle, for which a relatively large number of orthologous genes are available and for which mice and rats can be used as outgroups (see Discussion). In this study we used the paleontological estimates of the divergence times between orangutans and humans (about 13 MY) and that between primates and artiodactyls (about 90 MY) as the calibration points. The fossil dating of 13 MY for the human–orangutan divergence was once questioned (Pilbeam *et al.* 1990), but recent study suggests that "the *Sivapithecus–Pongo* clade remains the strongest phylogenetic hypothesis" and thus 13 MY is reasonable (Ward 1997). Archibald (1996) and Archibald *et al.* (2001) reported fossil remains of archaic ungulates that suggest that the divergence between ungulates and other orders of placental mammals occurred 85–90 MY ago. In this article we assume that the artiodactyl and primate lineages diverged 90 MY ago, because the fossil record usually gives underestimates of divergence times. We use the two calibration points to examine whether each of the calibration points gives similar time estimates for other divergence times. The number of orthologous proteins used was 27, with a total of 6586 amino acids (http://www.bio.psu.edu/people/faculty/nei/lab).

The phylogenetic tree with least-squares estimates of branch lengths is presented in Figure 3A. Takezaki *et al.*'s *U* statistic showed that the molecular clock hypothe-

Table 2. Estimates (± standard errors) of divergence times of the human lineage from other primate species and artiodactyls obtained using Poisson correction (PC) gamma distances

Calibration point	Chimp (t_1)	Gorilla (t_2)	Orangutan (t_3)	Artiodactyl (t_4)
Concatenated distance (CD) approach				
Multiprotein gamma (d_{MG}; $a = 0.54$)				
$t_4 = 90$ MY	5.7 ± 1.5	6.0 ± 1.4	12.0 ± 2.6	90
$t_3 = 13$ MY	6.2 ± 0.9	6.6 ± 0.8	13	97.8 ± 22.7
Weighted average (d_{WA}; $\bar{a} = 0.95$)				
$t_4 = 90$ MY	5.4 ± 1.5	5.7 ± 1.3	11.3 ± 2.6	90
$t_3 = 13$ MY	6.2 ± 0.9	6.5 ± 0.7	13	103.3 ± 26.8
Dayhoff PAM distance ($a = 2.25$)				
$t_4 = 90$ MY	6.6 ± 1.7	7.0 ± 1.5	13.7 ± 2.4	90
$t_3 = 13$ MY	6.3 ± 0.8	6.6 ± 0.7	13	85.6 ± 17.5
Individual protein (IP) approach ($\bar{a} = 0.95$)				
$t_4 = 90$ MY	4.4 ± 0.8	6.8 ± 2.0	12.1 ± 1.3	90
$t_3 = 13$ MY	5.3 ± 0.9	7.0 ± 1.1	13	143.7 ± 24.0

Twenty-seven nuclear genes with a total of 6586 amino acids were used.

sis cannot be rejected at the 1% level. We therefore constructed a linearized tree and estimated the times of divergence for the four branch points involved (t_1, t_2, t_3, and t_4 in Figure 3B). When multiprotein gamma distance was used with the calibration point of $t_4 = 90$ MY (primate–artiodactyl divergence), we obtained 5.7 MY for the divergence between humans and chimpanzees (t_1), 6.0 MY between humans and gorillas (t_2), and 12 MY between humans and orangutans (t_3) (Table 2). The estimate of t_3 is a little smaller than the fossil dating (13 MY), but the difference is not statistically significant. By contrast, if we use $t_3 = 13$ MY as the calibration point, we obtain $\hat{t}_1 = 6.2$, $\hat{t}_2 = 6.6$, and $\hat{t}_4 = 97.8$ MY. These estimates are very close to those obtained using $t_4 = 90$ MY as the calibration point, if we consider the extent of standard errors attached. The estimate of t_4 is also rather close to the calibration point of 90 MY. Essentially the same results are obtained when weighted average distances (d_{WA}) with an estimated average a value of 0.95 are used, though the use of calibration point $t_4 = 90$ MY tends to give smaller estimates of other divergence times and the use of $t_3 = 13$ MY gives a rather high estimate of t_4.

Since there is a possibility that the a value obtained from closely related sequences is an underestimate, we also tried Dayhoff's PAM distance (actually d_{MG} with $a = 2.25$) to estimate divergence times. This distance gave more or less the same results as those obtained by d_{MG} with $a = 0.54$. This indicates that variation in the a value does not seriously affect the time estimates in the present case as long as a is 0.54–2.25.

When the IP approach is used with calibration point $t_4 = 90$ MY, the estimates of t_2 and t_3 are similar to those obtained by d_{MG} in the CD approach, but the estimate of t_1 is 4.4 MY and is considerably smaller than those by the CD approach. This estimate is also smaller than the dating of 5.2–6.0 MY of the currently available fossil remains in the human lineage (Aiello and Collard 2001; Haile-Selassie 2001). By contrast, if we use $t_3 = 13$ MY as the calibration point, the estimate of t_4 becomes very high, though the standard error of the estimate is also very high.

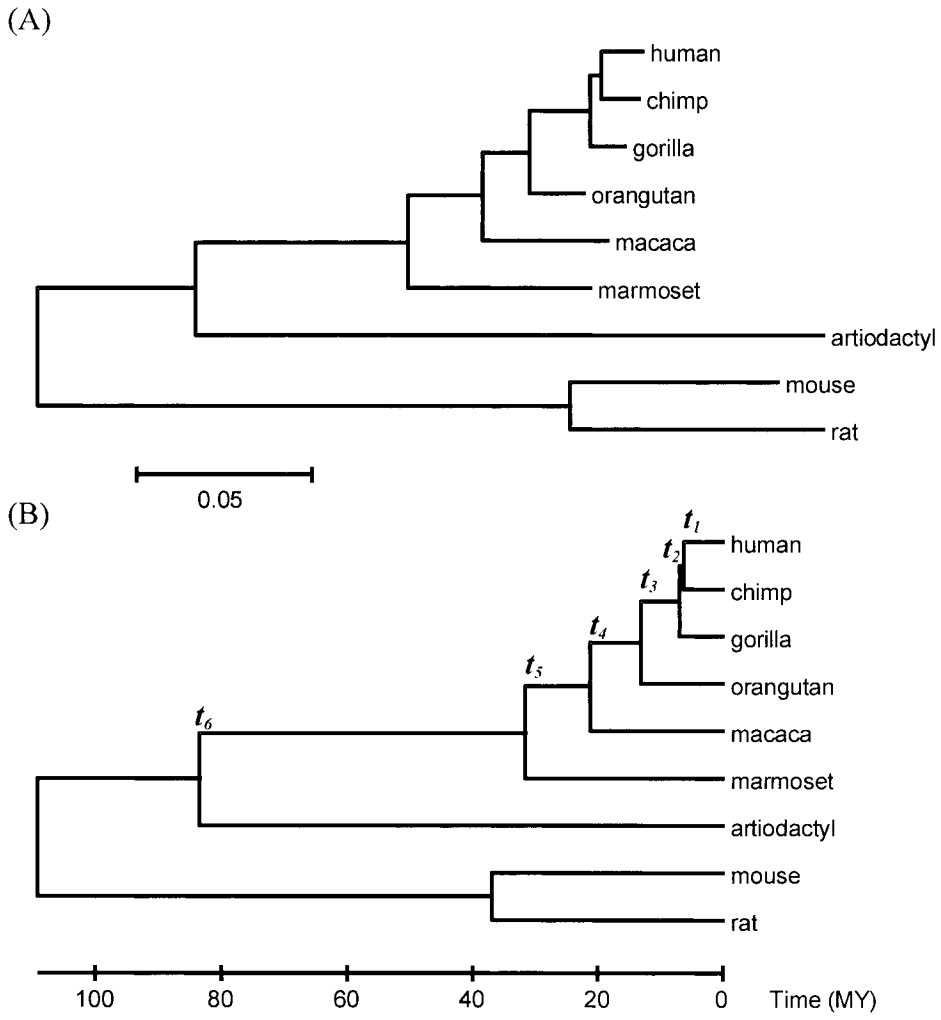


Figure 4. **(A)** NJ tree for primates constructed using multiprotein gamma distance (d_{MG}) with $a = 0.51$ for 12 protein sequences. **(B)** Linearized tree.

Six Species of Higher Primates

In the above study we did not include any species of OW and NW monkeys, because not many orthologous genes were available from these species. However, we could collect 12 shared nuclear proteins (2305 amino acids) from OW monkeys (macaque) and NW monkeys (mostly marmoset), as well as from several other species given in Figure 4. These genes were a subset of the genes used in the study of hominoid species (see http://www.bio.psu.edu/people/faculty/nei/lab). Although the number of genes available is quite small, we estimated the times of divergence of the OW and NW monkeys from the human lineages.

The phylogenetic tree for the six pri-

Table 3. Estimates of divergence times of the human lineage from other primate species and artiodactyls

Calibration point	Chimp (t_1)	Gorilla (t_2)	Orangutan (t_3)	OWM (t_4)	NWM (t_5)	Artiodactyl (t_6)
Concatenated distance (CD) approach						
Multiprotein gamma (d_{MG}; $a = 0.51$)						
$t_6 = 90$ MY	6.3 ± 3.9	6.8 ± 2.9	14.2 ± 4.7	23.5 ± 4.1	35.2 ± 1.4	90
$t_3 = 13$ MY	5.8 ± 2.5	6.3 ± 0.9	13	21.5 ± 5.8	32.3 ± 14.7	82.7 ± 39.3
Weighted average (d_{WA}; $\bar{a} = 0.82$)						
$t_6 = 90$ MY	7.8 ± 2.9	9.0 ± 2.6	14.5 ± 5.8	23.4 ± 5.3	34.8 ± 2.8	90
$t_3 = 13$ MY	7.0 ± 1.8	8.0 ± 1.3	13	20.9 ± 6.9	31.1 ± 18.1	80.4± 52.9
Dayhoff PAM distance ($a = 2.25$)						
$t_6 = 90$ MY	7.6 ± 4.8	8.2 ± 3.7	16.6 ± 5.9	26.9 ± 3.2	39.3 ± 2.2	90
$t_3 = 13$ MY	5.9 ± 2.6	6.4 ± 1.0	13	21.0 ± 5.8	30.7 ± 14.5	70.3 ± 36.4
Individual protein (IP) approach ($\bar{a} = 0.82$)						
$t_6 = 90$ MY	3.4 ± 1.3	6.7 ± 2.7	11.5 ± 2.2	23.6 ± 3.9	33.6 ± 3.5	90
$t_3 = 13$ MY	3.8 ± 1.2	7.7 ± 1.9	13	33.0 ± 6.4	55.4 ± 9.6	175.6 ± 48.2

Twelve nuclear proteins with a total of 2305 amino acids were used.
OWM: Old World monkeys; NWM: New World monkeys.

mate species, artiodactyls, and two rodent species is presented in Figure 4A. This tree was obtained by using the multiprotein gamma distance (d_{MG}) for 12 protein sequences, and rodents were used as outgroup species. In this tree the artiodactyl proteins evolved significantly faster than other nonrodent proteins at the 5% level but not at the 1% level. We therefore constructed a linearized tree for nonrodent species (Figure 4B) and estimated the times of separation of the human lineage from the remaining primate species (Table 3).

The time estimates for chimpanzees, gorillas, and orangutans when d_{MG} is used are roughly the same as those in Table 2. When d_{WA} is used, the estimates tend to be higher. The time estimates for OW and NW monkeys are 22–24 MY and 32–35 MY, respectively, and the two calibration points give similar estimates. The time estimates obtained for artiodactyls when $t_3 = 13$ MY is used as the calibration point are smaller than those in Table 2 by 10–20 MY, but this difference is small compared with the standard errors of the estimates. Dayhoff's PAM distance also gave similar time estimates when $t_6 = 90$ MY was used as the calibration point. However, when $t_3 = 13$ MY was used as the calibration point, the estimate of the primate–artiodactyl divergence time was rather small.

The estimates obtained by the IP approach were quite unreasonable. When $t_6 = 90$ MY was used as the calibration point, the estimates of t_1, t_2, and t_3 were unreasonably low, and when $t_6 = 13$ MY was used, $\hat{t}_5$ and $\hat{t}_6$ were very large. These unreasonable results were obtained partly because the IP method gives biased estimates and partly because the number of proteins used was small.

Discussion

Generally speaking, estimation of divergence times from molecular data is more difficult than inference of the topology of a phylogenetic tree, because most genes do not evolve at a constant rate. This is particularly so when the number of genes or proteins used is small. In Figure 2, we see that the estimates of t_1 and t_2 from a single gene vary extensively, so that the estimates are not reliable. Our preliminary study about the number of genes required for obtaining reasonably good estimates of t_1 and t_2 from the dataset used in Table 1 suggested that at least 20 genes should be used. However, this depends on the relationship between the calibration point and the time to be estimated. In general, we need more genes for estimating times that are older than the calibration point than for estimating times younger than the calibration point. This is obvious from Table 3, where the standard errors of $\hat{t}_4$ and $\hat{t}_5$ relative to the means when $t_3 = 13$ MY was used as the calibration point are greater than those when $t_4 = 90$ MY was used. Of course, the number of genes required also depends on the number of species used. When a large number of closely related species are used, the number can be relatively small because the standard error of a branch point in a tree generally decreases as the number of species increases (Takezaki *et al.* 1995).

However, as was mentioned by many authors in the past, the most serious error in time estimates is generated when inappropriate calibration points are used. Unfortunately there are not many reliable paleontological data that can be used for producing reliable timescales. In the case of primate species, we used two different calibration points and obtained similar estimates for other divergence times. This gives a higher credibility to our estimates than those obtained by a single calibration point.

The fossil record seems to be powerless in determining the branching order of primates, artiodactyls, and rodents (Benton 1993). Most molecular studies in the past have suggested that the primate and artiodactyl lineages are closer to each other than to the rodent. Recently, however, using 19 nuclear and 3 mitochondrial genes and a Bayesian method of phylogenetic inference, Murphy *et al.* (2001) presented a phylogenetic tree that suggests a closer relationship between primates and rodents than between primates and artiodactyls. Almost all interior branches (or clades) of this tree are supported by a high Bayesian posterior probability. However, since our unpublished work (Suzuki Y and Nei M) indicates that the posterior probability of a Bayesian tree often gives an overestimate of statistical confidence, we are not convinced that the Murphy tree is well established. In fact, when we studied the phylogenetic relationships for humans, artiodactyls, and rodents by using 71 protein sequences with a total of 24,952 amino acids (Glazko GV and Nei M, unpublished), we obtained the tree presented in Figure 5. Here, chicken and *Xenopus* were used as outgroups. This tree shows quite clearly that humans are closer to artiodactyls than to rodents. The Bayesian phylogenetic inference (Huelsenbeck and Ronquist 2001) for the same dataset also gave the same topology and a credibility value (posterior probability) of 100% for all interior branches. The closer relationship between humans and artiodactyls than between humans and rodents was also supported by an analysis of another set of 20 proteins (Misawa K and Nei M, unpublished data). For these reasons, we decided to use rodents as the outgroup species in this study, though this has to be confirmed by using more genes and more species.

However, it should be noted that our estimates of divergence times for primate species do not depend so much on the phylogenetic position of rodents, because we used a linearized tree that was constructed by disregarding rodent species. Therefore, whether rodents are close to

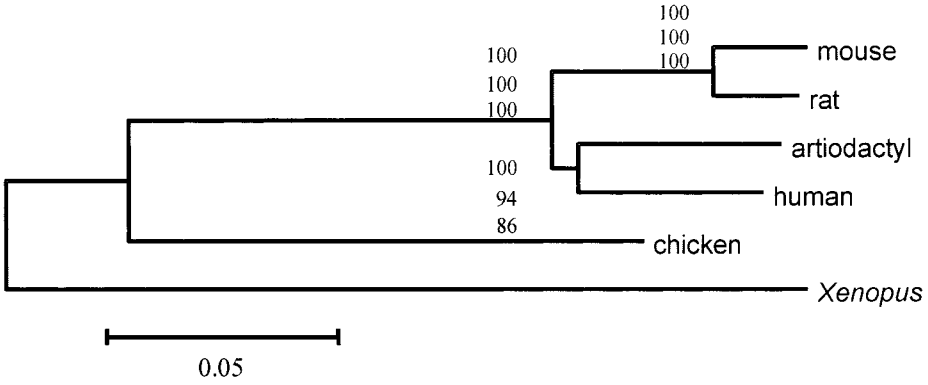


Figure 5. Phylogenetic tree for primates (humans), artiodactyls (mainly cattle genes), and rodents (mouse and rat) when chicken and *Xenopus* are used as outgroups. The tree represents an NJ tree obtained with PC distance for 71 protein sequences, but the MP and ML trees showed the same topology. The uppermost, middle, and bottom numbers given for each interior branch (or clade) are the bootstrap values for the NJ, MP, and ML trees, respectively. The bootstrap values for the ML tree are approximate because we used the RELL algorithm of Adachi and Hasegawa's (1996) MOLPHY program package. The Bayesian posterior probability was 100% for every interior branch when a Bayesian tree was constructed. The Bayesian tree was constructed by using the computer program MRBAYES (Huelsenbeck and Ronquist 2001). Four chains were run for 300,000 generations with a temperature parameter value of 0.2. Trees were sampled every 20 generations from the last 150,000 generations, and a total of 7500 sampled trees were used for inferring the Bayesian tree.

primates or artiodactyls, we will have more or less the same time estimates.

Some authors have attempted to estimate divergence times by using several local evolutionary rates (e.g., Yoder and Yang 2000). This approach gives various estimates of divergence time depending on the number of different local rates assumed, and at this stage it is difficult to choose the most appropriate ones. We should realize that reliable calibration points are generally scarce and the time estimates obtainable are quite crude. Therefore we had better not make too many assumptions. Of course, if different groups of organisms are known to evolve at different rates, divergence times should be estimated separately for different groups.

It should also be noted that the evolutionary rate varies with the type of gene used. Therefore it seems to be important to use the same type of gene in a given study. Some authors (e.g., Murphy *et al.* 2001; Wray *et al.* 1996) used a mixture of nuclear genes and mtDNA genes. Since evolutionary rate varies with the gene and the gamma parameter value varies with the type of gene, it is advisable to use the same type of gene as much as possible. In particular, mtDNA genes are known to evolve so erratically that they do not appear to be suitable for estimating divergence times (Glazko GV and Nei M, unpublished data).

References

Adachi J and Hasegawa M, 1996. MOLPHY: programs for molecular phylogenetics. Tokyo, Japan: Institute of Statistical Mathematics.

Aiello LC and Collard M, 2001. Palaeoanthropology: our newest oldest ancestor? Nature 410:526–527.

Archibald JD, 1996. Fossil evidence for a Late Cretaceous origin of "hoofed" mammals. Science 272:1150–1153.

Archibald JD, Averianov AO, and Ekdale EG, 2001. Late Cretaceous relatives of rabbits, rodents, and other extant eutherian mammals. Nature 414:62–65.

Arnason U, Gullberg A, and Janke A, 1998. Molecular timing of primate divergences as estimated by two nonprimate calibration points. J Mol Evol 47:718–727.

Arnason U, Gullberg A, Janke A, and Xu X, 1996. Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. J Mol Evol 43:650–661.

Ayala FJ, Rzhetsky A, and Ayala FJ, 1998. Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. Proc Natl Acad Sci USA 95:606–611.

Benton MJ, 1993. The fossil record 2. New York: Chapman & Hall.

Britten RJ, 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398.

Bromham L, Phillips MJ, and Penny D, 1999. Growing up with dinosaurs: molecular dates and the mammalian radiation. Trends Ecol Evol 14:113–118.

Cao Y, Fujiwara M, Nikaido M, Okada N, and Hasegawa M, 2000. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene 259:149–158.

Dayhoff MO, Schwartz RM, and Orcutt BC, 1987. A model of evolutionary change in proteins. In: Atlas of protein sequence and structure (Dayhoff MO, ed). Silver Springs, MD: National Biomedical Research Foundation; 345–352.

Dickerson RE, 1971. The structure of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26–45.

Doolittle RF, Feng D-F, Tsang S, Cho G, and Little E, 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477.

Duret L, Gouy M, and Mouchiroud D, 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res 22:2360–2365.

Easteal S, 1999. Molecular evidence for the early divergence of placental mammals. BioEssays 21:1052–1059.

Eizirik E, Murphy WJ, and O'Brien SJ, 2001. Molecular dating and biogeography of the early placental mammal radiation. J Hered 92:212–219.

Feng DF, Cho G, and Doolittle RF, 1997. Determining divergence times with a protein clock: update and reevaluation. Proc Natl Acad Sci USA 94:13028–13033.

Foote M, Hunter JP, Janis CM, and Sepkoski JJ Jr, 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. Science 283:1310–1314.

Gissi C, Reyes A, Pesole G, and Saccone C, 2000. Lineage-specific evolutionary rate in mammalian mtDNA. Mol Biol Evol 17:1022–1031.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, and Groves CP, 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9:585–598.

Grishin NV, 1995. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 41:675–679.

Gu X, 1998. Early metazoan divergence was about 830 million years ago. J Mol Evol 47:369–371.

Gu X and Li W-H, 1992. Higher rates of amino acid substitution in rodents than in humans. Mol Phylogenet Evol 1:211–214.

Gu X and Zhang J, 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol Biol Evol 15:1106–1113.

Haile-Selassie Y, 2001. Late Miocene hominids from the Middle Awash, Ethiopia. Nature 412:178–181.

Horai S, Hayasaka K, Kondo R, Tsugane K, and Takahata N, 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92:532–536.

Huelsenbeck JP and Ronquist F, 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.

Kumar S and Hedges SB, 1998. A molecular timescale for vertebrate evolution. Nature 392:917–919.

Li W-H, Gouy M, Sharp PM, O'hUigin C, and Yang YW, 1990. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. Proc Natl Acad Sci USA 87:6703–6707.

Margoliash E, 1963. Primary structure and evolution of cytochrome *c.* Proc Natl Acad Sci USA 50:672–679.

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder O, Stanhope MJ, de Jong WW, and Springer MS, 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.

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

Nei M, Chakraborty R, and Fuerst PA, 1976. Infinite allele model with varying mutation rate. Proc Natl Acad Sci 73:4164–4168.

Nei M and Kumar S, 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nei M, Xu P, and Glazko G, 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc Natl Acad Sci USA 98:2497–2502.

O'hUigin C and Li W-H, 1992. The molecular clock ticks regularly in muroid rodents and hamsters. J Mol Evol 35:377–384.

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

Pilbeam D, Rose MD, Barry JC, and Shah SM, 1990. New *Sivapithecus humeri* from Pakistan and the relationship of *Sivapithecus* and *Pongo.* Nature 348:237–239.

Saitou N and Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Su C and Nei M, 1999. Fifty-million-year-old polymorphism at an immunoglobulin variable region gene locus

in the rabbit evolutionary lineage. Proc Natl Acad Sci USA 96:9710–9715.

Takahashi K, Rooney AP, and Nei M, 2000. Origins and divergence times of mammalian class II MHC gene clusters. J Hered 19:189–204.

Takahata N and Satta Y, 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci USA 94:4811–4815.

Takezaki N, Rzhetsky A, and Nei M, 1995. Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823–833.

Waddell PJ, Kishino H, and Ota R, 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Informatics 12:141–154.

Ward S, 1997. The taxonomy and phylogenetic relationships of *Sivapithecus* revisited. In: Function, phylogeny and fossils: Miocene hominoid evolution and adaptation (Begun DR, Ward CV, and Rose MD, eds). New York: Plenum; 269–290.

Wilson AC, Carlson SS, and White TJ, 1977. Biochemical evolution. Annu Rev Biochem 46:573–639.

Wray GA, Levinton JS, and Shapiro LH, 1996. Molecular evidence for deep Precambrian divergences among metazoan phyla. Science 274:568–573.

Yoder AD and Yang Z, 2000. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17:1081–1090.

Zhang J and Gu X, 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics 149:1615–1625.

Zuckerkandl E and Pauling L, 1962. Molecular disease, evolution, and genetic heterogeneity. In: Horizons in biochemistry (Kasha M and Pullman B, eds). New York: Academic Press; 189–225.

Corresponding Editor: Oliver Ryder