An Analysis of the Dynamics of Mammalian Mitochondrial DNA Sequence Evolution[1]

Charles F. Aquadro, Norman Kaplan,† and Kenneth J. Risko†*

*Laboratory of Genetics and †Biometry and Risk Assessment Program,
National Institute of Environmental Health Sciences

and the model's predictions have been compared with estimates obtained from recent mtDNA sequence data for an increasingly divergent series of primates, the mouse and the cow (Anderson et al. 1981, 1982; Bibb et al. 1981; Brown et al. 1982). The results are consistent with the hypothesis that the decrease in the proportion of transitions observed as divergence increases is a consequence of the highly biased substitution process. In addition, the results support the hypothesis that, although a portion of the mtDNA molecule evolves at an extremely rapid rate, a significant portion of the molecule is under strong selective constraints.

Introduction

The mitochondrial DNA (mtDNA) genome of animals is rapidly becoming one of the best-characterized regions of the eucaryotic genome. The complete molecule (approximately 16,500 base pairs [bp] in mammals) has been sequenced for an individual human, house mouse, and cow (Anderson et al. 1981, 1982; Bibb et al. 1981). Portions of the molecule have also been sequenced for six additional humans (Walberg and Clayton 1981; Greenberg et al. 1983), several individual rats (e.g., Sekiya et al. 1980; Brown et al. 1981; Grosskopf and Feldmann 1981*a,* 1981*b;* Kobayashi et al. 1981; Saccone et al. 1981; Brown and Simpson 1982; Koike et al. 1982; Wolstenholme et al. 1982), and a series of increasingly divergent primates (Brown et al. 1982).

Comparisons of these mtDNA sequences have revealed several fascinating new features of sequence variation and evolution and have confirmed others which were predicted by earlier heteroduplex DNA melting and restriction map studies of mtDNA from a variety of organisms (e.g., Upholt and Dawid 1977; Brown et al. 1979; Avise et al. 1979; see Avise and Lansman [1983] and Brown [1983] for recent reviews). Some of the more interesting findings are that (1) among vertebrates, and particularly mammals, there is a surprising conservation of the arrangement of genetic material; (2) although certain regions of the mtDNA genome evolve faster than others, there are no large conserved blocks; (3) mtDNA diverges very rapidly among closely related individuals and species, yet the amount of divergence appears to plateau at approximately 30% after 20–40 million years (Myr); (4) substitution pathways are strongly biased in favor of transitions (purine [A, G] to purine or pyrimidine [C, T] to pyrimidine substitutions) in comparisons among individuals within a species, yet comparisons

[1] Key words: mitochondrial DNA, sequence evolution, base substitutions, transitions and transversions, mathematical model.

Address for correspondence and reprints: Charles F. Aquadro, Laboratory of Genetics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709.

Mol. Biol. Evol. 1(5):423–434. 1984.

of divergent mtDNA sequences (e.g., between mouse and cow) show a much smaller bias favoring transitions; and (5) this transition bias is seen in noncoding regions as well as regions coding for proteins, ribosomal RNAs, and transfer RNAs.

Several mathematical models of the substitution process at the nucleotide level have been proposed for the purpose of correcting for multiple substitutions at the same nucleotide site when estimating the rate of substitution per nucleotide site (see Brown et al. [1982] and references in Kaplan [1983]). Since mtDNA sequence data are now available for a series of species that have sufficiently different divergence times, it is possible to determine how well these models predict the temporal behavior of quantities which describe aspects of the dynamics of the substitution process. In this paper we study the temporal behavior of the predicted values of several quantities and examine how well these predictions agree with estimates obtained from recent mammalian mtDNA sequence data.

Analysis

Basic Model

Changes at the nucleotide level can occur by base pair substitution, insertion of nucleotides, deletion of nucleotides, or structural rearrangement. For the latter three phenomena, no reasonable mathematical models have been developed, and so they will not be considered. In addition, except for the D-loop region, they play a relatively minor role in the evolution of mammalian mtDNA (Brown 1983).

The substitutional process for a particular nucleotide site is usually modeled by a continuous-time, four-state Markov chain (see Kaplan [1983] and references therein). For mathematical convenience it will be assumed that a nucleotide site is in state 1, 2, 3, or 4 if it is an A, G, C, or T, respectively. This labeling differs from that which is used customarily.

Let q_{ij} denote the instantaneous rate of change from state i to state j, $i \neq j$, and set $q_{ii} = -\Sigma_{j \neq i}\, q_{ij}$. The q_{ij} do not change over time, and so the rates of base pair substitution are constant. It follows from the theory of Markov processes (Karlin 1969) that the matrix $Q = (q_{ij})_{1 \leq i,j \leq 4}$ contains all the information needed to construct the process except the initial frequencies of the four states (i.e., the ancestral base composition). For the purposes of this paper, these frequencies are assumed to be the stationary ones. Let $\mathbf{p} = (p_1, p_2, p_3, p_4)$ denote the stationary probabilities. It is known (Karlin 1969) that $\mathbf{p}$ satisfies

$$\mathbf{p}Q = \mathbf{0}. \tag{1}$$

Since many of the models that have been proposed in the literature are special cases of Kimura's (1981) six-parameter model, it will be assumed in this paper that Q is of the form

$$Q = \begin{bmatrix} -(\alpha_1 + 2\alpha) & \alpha_1 & \alpha & \alpha \\ \alpha_2 & -(\alpha_2 + 2\alpha) & \alpha & \alpha \\ \beta & \beta & -(\beta_1 + 2\beta) & \beta_1 \\ \beta & \beta & \beta_2 & -(\beta_2 + 2\beta) \end{bmatrix},$$

where α, α_1, α_2, β, β_1, and β_2 are positive parameters. The number of parameters cannot be reduced by assuming that $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ since this would imply that

$p_1 = p_2$ and $p_3 = p_4$. These latter equalities, however, are not consistent with the data (Brown 1983).

Suppose two homologous sequences diverged t years ago. We define $h_{ij}(t)$ = probability that a nucleotide site is in state i in the first sequence and state j in the second sequence; $f_{ij}(t) = h_{ij}(t) + h_{ji}(t)$; $D(t)$ = probability that a nucleotide site is in different states in the two sequences; and $T(t)$ = probability that a nucleotide site is in the transitional state, given that it is in different states in the two sequences. A nucleotide site is said to be in the transitional state if the nucleotide site is in different states in the two sequences and the two states of the nucleotide site belong to the sets $\{1, 2\}$ (purines) or $\{3, 4\}$ (pyrimidines). In view of the definitions of $D(t)$ and $T(t)$, their trajectories will be called the divergence and transition curves.

It follows from the properties of the Markov process that $f_{ij}(t)$, $D(t)$, and $T(t)$ satisfy the following relationships:

$$f_{ij}(t) = 2 \sum_{l=1}^{4} p_l P_{li}(t) P_{lj}(t), \tag{2}$$

$$D(t) = \sum_{i \neq j} f_{ij}(t), \tag{3}$$

and

$$T(t) = \frac{[f_{12}(t) + f_{34}(t)]}{D(t)}, \tag{4}$$

where $P_{ij}(t)$ is the probability that t years after divergence a nucleotide site is in state j, given that it was initially in state i. To compute $f_{ij}(t)$, $D(t)$, and $T(t)$ for any particular set of parameter values, it is sufficient to compute the $P_{ij}(t)$. Details of how the $P_{ij}(t)$ are computed are given in the Appendix.

Since $f_{ij}(t) \to 2p_i p_j$ as $t \to \infty$, one implication of the formulas for $D(t)$ and $T(t)$ is that, for large values of t,

$$D(t) \sim 1 - \sum_{i=1}^{4} p_i^2 \tag{5}$$

and

$$T(t) \sim \frac{2(p_1 p_2 + p_3 p_4)}{1 - \sum_{i=1}^{4} p_i^2}. \tag{6}$$

Thus, the asymptotic values of $D(t)$ and $T(t)$ can be calculated when only the p_i are known. The expressions on the right in equations (5) and (6) will be denoted by $D(\infty)$ and $T(\infty)$, respectively.

Data

Recently, Brown et al. (1982) sequenced a homologous mtDNA segment of approximately 896 bp from five different mammalian species: human, chimpanzee, gorilla, orangutan, and gibbon. This region, which contains the genes for three transfer RNAs and parts of two proteins, had previously been sequenced in another human

(Anderson et al. 1981), a cow (Anderson et al. 1982), and a mouse (Bibb et al. 1981). Because of insertions and deletions, the region was not the same length in all species. We have deleted from the analysis the eight nucleotide sites involved in this length variation, and we focus on the remaining 892-bp segment.

There is independent evidence (Sarich and Wilson 1967; Brown et al. 1979) suggesting that chimpanzee, gorilla, and human diverged about 5 Myr ago; orangutan and human 8 Myr ago; gibbon and other primates 10 Myr ago; and the primates, mouse, and cow 80 Myr ago. Hence, one can obtain estimates of $f_{ij}(t)$, $D(t)$, and $T(t)$ from these mtDNA sequence data for four different divergence times: 5, 8, 10, and 80 Myr. These estimates are given in table 1. The estimates of $f_{ij}(t)$, $D(t)$, and $T(t)$ from the comparison of the two human mtDNA sequences were not included in our analysis since only two nucleotides differed between these sequences (both differences were, however, transitions). The estimate of $T(t)$ at this low level of divergence is of limited use since it can only take on the values 0, 0.5, or 1.0.

Divergence Dynamics

Several reasonable estimates of the p_i are given in table 2. The associated predictions of $D(\infty)$ and $T(\infty)$, which are also given in table 2, are in the vicinity of

Table 1
Estimates and Adjusted Predicted Values of $f_{ij}(t)$, $D(t)$, and $T(t)$ for Mammalian mtDNA

	t (Time since Divergence in Myr)			
Variable	5	8	10	80
$f_{12}(t)$	(.0235, .0291)	(.0381, .0415)	(.0364, .0448)	(.0258, .0404)
	.0265	.0398	.0415	.0336
	.0557	.0649	.0674	.0459
$f_{13}(t)$	(.0022, .0078)	(.0224, .0224)	(.0269, .0325)	(.0796, .1233)
	.0056	.0024	.0289	.0946
	.0073	.0117	.0146	.0771
$f_{14}(t)$	(.0011, .0045)	(.0112, .0112)	(.0112, .0157)	(.0359, .0874)
	.0022	.0112	.0135	.0545
	.0067	.0107	.0132	.0657
$f_{23}(t)$	(.0000, .0000)	(.0011, .0045)	(.0056, .0090)	(.0045, .0112)
	.0000	.0036	.0074	.0088
	.0037	.0056	.0067	.0258
$f_{24}(t)$	(.0000, .0000)	(.0000, .0022)	(.0011, .0022)	(.0056, .0112)
	.0000	.0011	.0020	.0087
	.0033	.0049	.0059	.0219
$f_{34}(t)$	(.0583, .0695)	(.0796, .0942)	(.0874, .0998)	(.0830, .1233)
	.0632	.0849	.0926	.1102
	.0556	.0765	.0867	.1010
$D(t)$	(.087, .107)	(.158, .170)	(.182, .190)	(.269, .339)
	.098	.163	.186	.311
	.133	.174	.194	.337
$T(t)$	(.905, .937)	(.752, .776)	(.687, .740)	(.390, .532)
	.921	.764	.721	.465
	.841	.811	.792	.435

NOTE.—For any two sequences, $\hat{f}_{ij}(t)$ = (number of nucleotide sites in state i in either sequence and state j in the other)/892, $\hat{D}(t)$ = (number of nucleotide sites in different states)/892, and $\hat{T}(t)$ = (number of nucleotide sites in transitional state)/(number of nucleotide sites in different states). For each variable, the upper entry in the table is the minimum and maximum of the pairwise estimates for those sequences having the appropriate divergence time. The middle entry is the average of these estimates, and the lower entry is the adjusted predicted value computed from the model having the Q matrix in fig. 2a. The value of g used to adjust the predicted values of $f_{ij}(t)$ and $D(t)$ is 0.48.

Table 2
Estimates of Stationary Frequencies of the Four Bases

mtDNA Region	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{D}(\infty)$	$\hat{T}(\infty)$
Human, mouse and cow:						
entire genomes	.329 (.31, .33)	.130 (.13, .14)	.272 (.27, .31)	.269 (.25, .27)	.73	.32
892-bp segment:						
Whole region, eight sequences	.321 (.30, .36)	.104 (.10, .11)	.312 (.25, .35)	.263 (.24, .29)	.72	.32
Variable sites, eight sequences (448 bp)	.300 (.27, .38)	.071 (.05, .09)	.397 (.27, .47)	.231 (.18, .29)	.69	.33

NOTE.—Presented are the estimates of the p_i, ($\hat{p}_i$), for the L-strands of the indicated sequences. The estimates of $D(\infty)$, [($\hat{D}(\infty)$], and $T(\infty)$, [$\hat{T}(\infty)$], are also given as calculated from eqq. (5) and (6), respectively. The $\hat{p}_i$ were computed as follows: $\hat{p}_i = (\sum_{j=1}^{n} r_{ij})/n$, where r_{ij} is the fraction of sites in the jth sequence that are in state i, and n is the number of sequences analyzed. The values in the parentheses in the table are the smallest and largest of the r_{ij}.

0.72 and 0.32, respectively. The estimate of $D(t)$ at 80 Myr is 0.31 (table 1), suggesting that the divergence curve continues to increase after 80 Myr to a value more than twice its value at 80 Myr. This large amount of additional divergence is surprising, especially since comparisons of mouse and *Drosophila yakuba* for two other coding regions show only 57% and 49% divergence after 600 Myr (Clary et al. 1982).

A possible explanation for the additional predicted divergence is that the estimates of the p_i are inaccurate. However, the plot in figure 1 shows that for the prediction of $D(\infty)$ to be smaller than 57%, for example, it is necessary that one of the p_i be greater than 0.47. Since this extreme condition is not approached in any extant

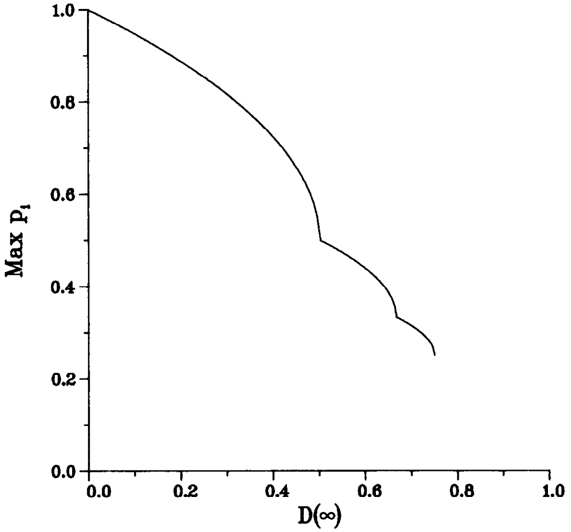

FIG. 1.—One of the p_i must be at least as big as max p_i in order to achieve a given value of $D(\infty)$. The relationship between max p_i and $D(\infty)$ is max $p_i = \frac{1}{2} + \{\frac{1}{2}[\frac{1}{2} - D(\infty)]\}^{1/2}$ if $0 \leq D(\infty) \leq \frac{1}{2}$; $\frac{1}{3} + \{\frac{1}{6}[\frac{2}{3} - D(\infty)]\}^{1/2}$ if $\frac{1}{2} \leq D(\infty) \leq \frac{2}{3}$; $\frac{1}{4} + \{\frac{1}{12}[\frac{3}{4} - D(\infty)]\}^{1/2}$ if $\frac{2}{3} \leq D(\infty) \leq \frac{3}{4}$. It follows from the definition of $D(\infty)$ (eq. [5]) that it is always less than or equal to 0.75.

mammalian mtDNA sequence, errors in the estimation of the p_i seem an unlikely explanation for the large prediction of $D(\infty)$.

An alternative hypothesis to account for the difference between the estimate of $D(t)$ at 80 Myr and the prediction of $D(\infty)$ is that only a portion of the region compared is free to change, whereas the remaining sites are under strong selective constraints. If g denotes the fraction of the 892 sites that are not free to undergo substitution (invariable sites), then the number of sites that can change (variable sites) is $892(1 - g)$. It follows that the denominator of the estimates of $D(t)$ and $f_{ij}(t)$ is too large. Thus, the estimates need to be inflated by the factor $1/(1 - g)$. Estimates of $T(t)$ do not have to be adjusted since the number of sites that cannot change does not enter into their computation. Rather than manipulate the data, we have chosen to adjust the predicted values of $D(t)$ and $f_{ij}(t)$ by multiplying them by the factor $1 - g$.

The number of nucleotide sites that are in the same state in all eight sequences (unvaried sites) is the sum of the number of invariable sites ($892g$) and the number of variable sites that are identical in the eight sequences by chance. Since the observed number of nucleotide sites that are unvaried is 444, an upper bound on g for these data is 0.497 (444/892). If the number of unvaried sites that occur by chance is small, then the value of g should be in the vicinity of 0.497.

Six parameters must be specified in order to compute the $f_{ij}(t)$, $D(t)$, and $T(t)$ for finite time points. An essential restriction on Q is that it satisfies equation (1). Thus if $\hat{\mathbf{p}}$ is an estimate of $\mathbf{p}$, it is reasonable to choose Q so that $\hat{\mathbf{p}}Q = \mathbf{0}$. Since there is no way to identify those nucleotide sites that are unvaried by chance, the estimates of the p_i, unless otherwise stated, are based on the entire 892-bp fragment (table 2). The constraint $\hat{\mathbf{p}}Q = \mathbf{0}$ leads to three equations, leaving three parameters to be determined. The subset of the parameter space to which these parameters were restricted was determined in the following way. It was not clear a priori how changes in the elements of Q affect $f_{ij}(t)$ and $D(t)$. Hence, it was more convenient to vary the initial slope of the divergence curve, $D'(0)$, and the initial value of the transition curve, $T(0)$. The values of $D'(0)$ and $T(0)$ are related to the elements of Q as follows:

$$D'(0) = -2 \sum_{i=1}^{4} p_i q_{ii}$$

and

$$T(0) = \frac{(p_1 q_{12} + p_2 q_{21} + p_3 q_{34} + p_4 q_{43})}{-\sum_{i=1}^{4} p_i q_{ii}} .$$

The value of $D'(0)$ was varied from $0.01/(1 - g)$ to $0.06/(1 - g)$ in increments of 0.01, and $T(0)$ was varied from 0.85 to 0.99 in increments of 0.01. These values were chosen to bracket those considered realistic for mammalian mtDNA (Brown et al. 1982; Aquadro and Greenberg 1983; Brown 1983). Finally, for each value of $D'(0)$ and $T(0)$, the remaining parameter (chosen for convenience to be α_1) was allowed to vary over all its possible values.

For each value of g ranging from 0.0 to 0.5 in increments of 0.1, and for each set of the six parameter values, the sum of squared differences (SSD) was computed between the estimates of $f_{ij}(t)$, $D(t)$, and $T(t)$ and the adjusted predicted values. In addition, a 99% confidence interval for the number of unvaried sites was determined (see Appendix).

Since the values of g associated with the 30 parameter sets having the smallest SSD were 0.4 or 0.5, the calculations above were repeated for values of g ranging from 0.4 to 0.5 in increments of 0.01. In figure 2*a*, the Q matrix is given for the model having the smallest SSD and whose 99% confidence interval for the number of unvaried sites (441.9, 467.8) contains 444. Plots of $(1 - g)D(t)$ and $T(t)$ corresponding to this Q matrix are presented in figure 2*b*, and the predicted values of the $(1 - g)f_{ij}(t)$ are given in table 1. The value of g for this model is 0.48, and the SSD is 0.0337. Four sets of parameter values led to slightly smaller SSDs (0.0318–0.0329), but their 99% confidence intervals did not contain 444. The values of g for these models were either 0.49 or 0.50. It is worth noting that for $g = 0.0$ the smallest SSD was 0.0587, and the associated 99% confidence interval was (336.9, 412.8). The predicted trajectories of $(1 - g)D(t)$ and $T(t)$ in figure 2*b* are reasonably close to the data, as are the predicted values of $(1 - g)f_{ij}(t)$ (table 1). The values of $(1 - g)D'(0)$ and $T(0)$ for the model are 0.038 and 0.89, respectively.

In calling nucleotide sites invariable, we refer only to the time frame of divergence over which we have sampled sequences (80 Myr). It seems reasonable to assume that

(a)

$$Q = \begin{pmatrix} -0.0356 & 0.0306 & 0.0025 & 0.0025 \\ 0.0997 & -0.1048 & 0.0025 & 0.0025 \\ 0.0019 & 0.0019 & -0.0268 & 0.0231 \\ 0.0019 & 0.0019 & 0.0278 & -0.0315 \end{pmatrix}$$

(b)

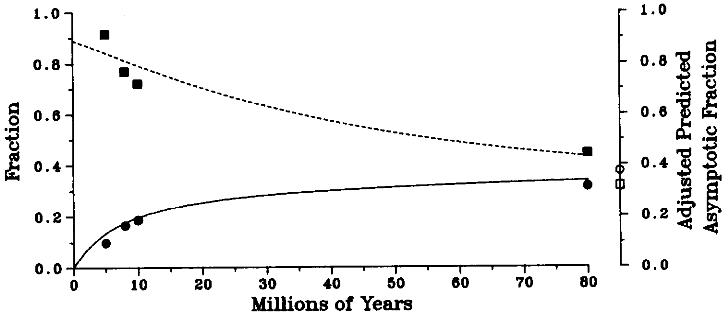

FIG. 2.—*a*, The Q matrix leading to the smallest SSD between the estimated and adjusted predicted values of $f_{ij}(t)$, $D(t)$, and $T(t)$. For this set of parameter values $g = 0.48$, $(1 - g)D'(0) = 0.02$, $T(0) = 0.89$, and SSD = 0.0337. *b*, The predicted divergence curve adjusted for invariant sites (——) and predicted transition curve (- - -) for the model having the Q matrix in fig. 2*a*. Estimates of $D(t)$ (●) and $T(t)$ (■) are obtained from table 1. The adjusted predicted value of $D(\infty)$ (○) and the predicted value of $T(\infty)$ (□) are plotted on the right-hand y-axis.

selective pressures may change over time, and that sites that were at one time invariable may be released from selective constraint due, for example, to substitutions at neighboring sites (in essence, the nucleotide analogy of the covarion hypothesis of Fitch and Markowitz [1970]). Likewise, sites we now see as invariable could have been variable at some distant time.

A distinguishing feature of the Q matrix in figure 2*a* is the strong transition bias, that is, when a substitution occurs, the probability is much larger that it will be a transition than a transversion. In fact, this transition bias was a property of every Q matrix examined whose 99% confidence interval contained the observed number of unvaried sites (444).

To examine the sensitivity of our results to different choices of the $\hat{p}_i$, the analysis above was repeated for the other estimates in table 2. Calculations not given here show that for this range of the $\hat{p}_i$ the SSD is not substantially reduced and that the models having the smallest SSD are qualitatively the same in their predictions and general features. A substantially poorer fit to the data, particularly the $f_{ij}(t)$, results if $p_1 = p_2 = p_3 = p_4 = 0.25$ as required by many simpler models (e.g., Jukes and Cantor 1969; Kimura 1980).

Two additional parameters were introduced to accommodate the observation that certain nucleotide sites evolve faster than others. It was assumed that a fraction, f, of the nucleotides are evolving L times as fast as the remaining nucleotides. The Q matrix for this model is $Q^* = (fL + 1 - f)Q$, where Q is the associated matrix for the usual model. Calculations not given here show that the SSD is not substantially reduced by enlarging the parameter space in this fashion.

The effect of errors in the estimates of the divergence times was also examined. The results are robust to errors in the largest divergence time (80 Myr) so long as it remains large relative to the other divergence times. The relative differences among the early times are generally believed to be correct (Andrews 1982; Andrews and Cronin 1982). However, their precise ages may be in error. Decreasing the divergence times acts to force up the rate of substitution and so makes it more difficult for unvaried variable sites to occur by chance. The resulting estimates of g are therefore very close, if not equal, to 444/892. In contrast, increasing the early divergence times decreases the rate of substitution and so makes it easier for unvaried variable sites to occur randomly. The estimates of g can thus be much smaller than 444/892. Nonetheless, the Q matrices associated with the models having a small SSD always show a transition bias.

Discussion

Our analysis supports the hypothesis that a substantial fraction (0.45–0.49) of the mammalian mtDNA segment examined is under strong selective constraints, and thus the rate of substitution at these sites is negligible. The nature of the constraints clearly varies over the mitochondrial genome. For example, constraints apparently related to their secondary structures exist for transfer RNAs. The striking uniformity of the location of conserved regions over the 22 mitochondrial transfer RNAs supports these secondary structure arguments (see esp. Anderson et al. 1982; Brown et al. 1982). Of the 195 nucleotides constituting the three transfer RNA genes in our comparison, 69% are conserved over the lineages leading to the six primates, mouse, and cow.

The remaining 697 nucleotides of the sequences examined in this paper represent portions of genes for two putative proteins. Among the primate, cow, and mouse sequences, 51% and 61% of the first and second codon positions, respectively, are conserved, whereas only 16% are conserved in the third position sites over all lineages. Presumably this pattern reflects the degeneracy of the genetic code and a higher rate of silent rather than replacement nucleotide substitutions (Bibb et al. 1981; Anderson et al. 1982; Brown and Simpson 1982; Brown et al. 1982; Miyata et al. 1982).

The apparent plateau of divergence at 30%–40% in mtDNA melting studies (see Brown et al. 1979) and direct comparisons of the complete sequences of human, mouse, and cow (Anderson et al. 1981, 1982; Bibb et al. 1981) suggest that the evolutionary dynamics observed for this 892-bp segment are reflective on average of the whole mammalian mitochondrial genome. Thus, it appears reasonable to extend the general conclusion of substantial constraints to the entire mammalian mitochondrial genome.

The Q matrix in figure 2*a* is representative of those parameter sets whose predictions (adjusted for invariable sites) are consistent with estimates of $f_{ij}(t)$, $D(t)$, and $T(t)$ (table 1 and fig. 2*b*). A notable consequence of the structure of this matrix is that, when a substitution occurs, it is far more likely to be a transition than a transversion. The temporal predictions of $T(t)$ for this model thus decrease in time, supporting the hypothesis (Brown et al. 1982; Holmquist 1983) that the decrease in the proportion of transitions observed in pairwise comparisons of increasingly divergent sequences is a consequence of the strong substitution bias favoring transitions and does not reflect temporal changes in the underlying substitution process over time. Inferences about the Q matrix based on sequences representing unknown times since divergence can therefore be misleading. For example, comparisons of partial sequences of mtDNAs from three related species of *Drosophila* reported by Clary et al. (1982) reveal that only 46% of the differences are transitions. Since we do not know how long these three sequences have been evolving independently from a common ancestor, it may be incorrect to conclude that there is not a major transitional bias in *Drosophila.* That is, if the three *Drosophila* species have been evolving as separate mtDNA lineages for a long period of time, any bias strongly favoring transitions may have eroded away in the observed divergence in the same way that it has in the primate comparisons of Brown et al. (1982). However, the low level of sequence divergence seen among the three sequences argues against this possibility (2.9% between *D. yakuba* and *D. melanogaster,* 8.4% and 8.6% between *D. virilis* and *D. yakuba* and *D. melanogaster,* respectively; Clary et al. [1982]). In contrast to the *Drosophila* data, the gibbon mtDNA sequence examined in this paper differs from that of human, chimp, gorilla, and orangutan at 9.8% of its nucleotides, yet 91.5% of these differences are transitions (Brown et al. 1982). More intraspecific mtDNA sequence comparisons are clearly needed to assess the phylogenetic distribution of the transition bias and to test the generality of our results.

Acknowledgments

We thank C. H. Langley, J. C. Stephens, B. S. Weir, M. Nei, and several anonymous reviewers for constructive criticism, and L. Gardner and S. Walker for their careful preparation of the manuscript.

APPENDIX

Calculations

Calculation of the $P_{ij}(t)$

Let $P(t)$ denote the matrix $[P_{ij}(t)]_{1\leq i,j\leq 4}$. It is well known (Karlin 1969) that

$$P(t) = \sum_{j=0}^{\infty} \frac{Q^j t^j}{j!}.$$

It follows from the spectral representation of Q (Karlin 1969) that there exist matrices A and B such that

$$P(t) = A \begin{pmatrix} e^{\lambda_1 t} & 0 & 0 & 0 \\ 0 & e^{\lambda_2 t} & 0 & 0 \\ 0 & 0 & e^{\lambda_3 t} & 0 \\ 0 & 0 & 0 & e^{\lambda_4 t} \end{pmatrix} B,$$

where the λ_i are the eigenvalues of Q. For any given set of parameters, the λ_i and the two matrices A and B were computed. The $P_{ij}(t)$ can also be calculated from equation 5 in Gojobori et al. (1982).

Calculation of the Confidence Interval for the Number of Unvaried Sites

Under the assumptions of the model, the number of unvaried variable sites has a binomial distribution with mean $892(1 - g)q$ and variance $892(1 - g)q(1 - q)$, where g is the fraction of sites that cannot change and q is the probability that a nucleotide site, which is capable of changing, is in the same state in all eight sequences. It thus follows from standard theory that a 99% confidence interval for the number of unvaried sites is $892[g + (1 - g)q] \pm 2.58[892(1 - g)q(1 - q)]^{1/2}$. In order to compute q it is necessary to know the branching order and time between successive branches. The formula for q for the sequences under consideration is

$$q = \sum_{j=1}^{4} \sum_{i=1}^{4} \sum_{i_1=1}^{4} \sum_{i_2=1}^{4} \sum_{i_3=1}^{4} \sum_{i_4=1}^{4} \times [p_i P_{ij}^2(80) P_{ii_1}(70) P_{i_1 j}(10) P_{i_1 i_2}(2) P_{i_2 j}(8) P_{i_2 i_3}(3) P_{i_3 j}^2(5) P_{i_3 i_4}(4.9) P_{i_4 j}^2(0.1)].$$

LITERATURE CITED

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. DE BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICK, B. A. ROE, F. SANGER, P. H. SCHREIER, A. J. H. SMITH, R. STADEN, and I. G. YOUNG. 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

ANDERSON, S., M. H. L. DE BRUIJN, A. R. COULSON, I. C. EPERON, F. SANGER, and I. G. YOUNG. 1982. Complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. J. Mol. Biol. **156**:683–717.

ANDREWS, P. 1982. Hominoid evolution. Nature **295**:185–186.

ANDREWS, P., and J. E. CRONIN. 1982. The relationship of *Sivapithecus* and *Ramapithecus* and the evolution of the orangutan. Nature **297**:541–546.

AQUADRO, C. F., and B. D. GREENBERG. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics **103**:287–312.

AVISE, J. C., and R. A. LANSMAN. 1983. Polymorphism of mitochondrial DNA in populations of higher animals. Pp. 147–164 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes. Sinauer, Sunderland, Mass.

AVISE, J. C., R. A. LANSMAN, and R. O. SHADE. 1979. The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *Peromyscus*. Genetics **92**:279–295.

BIBB, M. J., R. A. VAN ETTEN, C. T. WRIGHT, M. W. WALBERG, and D. A. CLAYTON. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell **26**:167–180.

BROWN, G. G., F. J. CASTORA, S. C. FRANTZ, and M. V. SIMPSON. 1981. Mitochondrial DNA polymorphism: evolutionary studies on the genus *Rattus*. Ann. N.Y. Acad. Sci. **361**:135–153.

BROWN, G. G., and M. V. SIMPSON. 1982. Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. Proc. Natl. Acad. Sci. **79**:3246–3250.

BROWN, W. M. 1983. Evolution of animal mitochondrial DNA. Pp. 62–88 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

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.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

CLARY, D. O., J. M. GODDARD, S. C. MARTIN, C. M.-R. FAURON, and D. R. WOLSTENHOLME. 1982. *Drosophila* mitochondrial DNA: a novel gene order. Nucleic Acids Res. **10**:6619–6637.

FITCH, W. M., and E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.

GOJOBORI, T., K. ISHII, and M. NEI. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. **18**:414–423.

GREENBERG, B. D., J. E. NEWBOLD, and A. SUGINO. 1983. Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene **21**:33–49.

GROSSKOPF, R., and H. FELDMANN. 1981*a*. Analysis of a DNA segment from rat liver mitochondria containing the genes for the cytochrome oxidase subunits I, II and III, ATPase subunit 6, and several tRNA genes. Curr. Genet. **4**:151–158.

———. 1981*b*. tRNA genes in rat liver mitochondrial DNA. Curr. Genet. **4**:191–196.

HOLMQUIST, R. 1983. Transitions and transversions in evolutionary descent: an approach to understanding. J. Mol. Evol. **19**:134–144.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KAPLAN, N. 1983. Statistical analysis of restriction enzyme map data and nucleotide sequence data. Pp. 75–106 *in* B. S. WEIR, ed. Statistical analysis of DNA sequence data. Dekker, New York.

KARLIN, S. 1969. A first course in stochastic processes. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

KOBAYASHI, M., T. SEKI, K. YAGINUMA, and K. KOIKE. 1981. Nucleotide sequences of small ribosomal RNA and adjacent transfer RNA genes in rat mitochondrial DNA. Gene **16**:297–307.

KOIKE, K., M. KOBAYASHI, K. YAGINUMA, M. TAIRA, E. YOSHIDA, and M. IMAI. 1982. Nucleotide sequence and evolution of the rat mitochondrial cytochrome b gene containing the ochre termination codon. Gene **20**:177–185.

MIYATA, T., H. HAYASHIDA, R. KIKUNO, M. HASEGAWA, M. KOBAYASHI, and K. KOIKE. 1982. Molecular clock of silent substitution: at least a six-fold preponderance of silent changes in mitochondrial genes over those in nuclear genes. J. Mol. Evol.. **19**:28–35.

SACCONE, C., P. CANTATORE, G. GADALETA, R. GALLERANI, C. LANAVE, G. PEPE, and A. M. KROON. 1981. The nucleotide sequence of the large ribosomal RNA gene and the adjacent *t*RNA genes from rat mitochondria. Nucleic Acids Res. **9**:4139–4148.

SARICH, V. M., and A. C. WILSON. 1967. Immunological time scale for hominid evolution. Science **158**:1200–1203.

SEKIYA, T., M. KOBAYASHI, T. SEKI, and K. KOIKE. 1980. Nucleotide sequence of a cloned fragment of rat mitochondrial DNA containing the replication origin. Gene **11**:53–62.

UPHOLT, W. B., and I. B. DAWID. 1977. Mapping of mitochondrial DNA of individual sheep and goats: rapid evolution in the D-loop region. Cell **11**:571–583.

WALBERG, M. W., and D. A. CLAYTON. 1981. Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic Acids Res. **9**:5411–5421.

WOLSTENHOLME, D. R., C. M.-R. FAURON, and J. M. GODDARD. 1982. Nucleotide sequence of *Rattus norvegicus* mitochondrial DNA that includes the genes for tRNA ile, tRNA gln and tRNA f-met. Gene **20**:63–69.

MASATOSHI NEI, reviewing editor

Received November 28, 1983; revision received March 5, 1984.