

HUMAN MITOCHONDRIAL DNA VARIATION AND EVOLUTION: ANALYSIS OF NUCLEOTIDE SEQUENCES FROM SEVEN INDIVIDUALS

CHARLES F. AQUADRO AND BARRY D. GREENBERG^{*1}

Laboratory of Genetics, National Institute of Environmental Health Sciences, Research Triangle
Park, North Carolina 27709, and ^{*}Curriculum in Genetics, University of North Carolina,
Chapel Hill, North Carolina 27514

Manuscript received April 19, 1982
Revised copy accepted October 20, 1982

ABSTRACT

We have analyzed nucleotide sequence variation in an approximately 900-base pair region of the human mitochondrial DNA molecule encompassing the heavy strand origin of replication and the D-loop. Our analysis has focused on nucleotide sequences available from seven humans. Average nucleotide diversity among the sequences is 1.7%, several-fold higher than estimates from restriction endonuclease site variation in mtDNA from these individuals and previously reported for other humans. This disparity is consistent with the rapidly evolving nature of this noncoding region. However, several instances of convergent or parallel gain and loss of restriction sites due to multiple substitutions were observed. In addition, other results suggest that restriction site (as well as pairwise sequence) comparisons may underestimate the total number of substitutions that have occurred since the divergence of two mtDNA sequences from a common ancestral sequence, even at low levels of divergence. This emphasizes the importance of recognizing the large standard errors associated with estimates of sequence variability, particularly when constructing phylogenies among closely related sequences. Analysis of the observed number and direction of substitutions revealed several significant biases, most notably a strand dependence of substitution type and a 32-fold bias favoring transitions over transversions. The results also revealed a significantly nonrandom distribution of nucleotide substitutions and sequence length variation. Significantly more multiple substitutions were observed than expected for these closely related sequences under the assumption of uniform rates of substitution. The bias for transitions has resulted in predominantly convergent or parallel changes among the observed multiple substitutions. There is no convincing evidence that recombination has contributed to the mtDNA sequence diversity we have observed.

DURING the last few years, DNA cloning and sequencing techniques have generated a wealth of nucleotide sequence data for a variety of genes and regions from a diverse array of organisms. Significant progress toward understanding the nature of molecular evolution has come from interspecific studies of nucleotide sequences of nuclear genes, as well as from comparisons among members of gene families such as the globins (reviewed by JEFFREYS 1981). The recent determinations of the complete mitochondrial DNA (mtDNA) sequences

¹ Present address: Department of Genetics, Stanford University, Stanford, CA 94305.

of man, cow and mouse have provided additional insights into sequence variation and evolution (ANDERSON *et al.* 1981, 1982; BIBB *et al.* 1981).

Our ability to absorb this new information into population and evolutionary genetic theory relies heavily on our recognition of the nature, frequency and distribution of sequence variation present within natural populations. The majority of the available comparisons, however, has been among single representatives of relatively old lineages, and it is possible that features of the substitution process have been significantly obscured by multiple base changes and recombination. No "population" sample of DNA sequences has been available with which to approach these questions. Recently, however, one of us has sequenced several overlapping stretches of mtDNA in a region surrounding the heavy strand origin of replication from five humans (GREENBERG, NEWBOLD and SUGINO 1982). Complete sequences for this region are also available for two additional humans (ANDERSON *et al.* 1981; WALBERG and CLAYTON 1981).

Mammalian mtDNA is a circular molecule, approximately 16,000 nucleotides in length, that is uniparentally inherited (e.g., HUTCHISON *et al.* 1974; GILES *et al.* 1980; BROWN *et al.* 1981; LANSMAN, AVISE and HUETTEL 1983) and apparently monoclonal within individuals (e.g., JAKOVCIĆ, CASEY and RABINOWITZ 1975; POTTER *et al.* 1975; AVISE *et al.* 1979; HAYASHI *et al.* 1979; however, see HAUSWIRTH and LAIPIS 1982). Recent models and simulations suggest that animal mtDNA sequence variation generated by mutation (e.g., nucleotide substitution, insertion, deletion, etc.) is rapidly fixed or lost within the germ cell lineage (UPHOLT and DAWID 1977; TAKAHATA and MARUYAMA 1981; BIRKY, MARUYAMA and FUERST 1982; CHAPMAN *et al.* 1982). In addition, the waiting time between mutations in a cell lineage appears to be orders of magnitude longer than the time to fixation of a new mutant (BIRKY, MARUYAMA and FUERST 1982). Thus, the probability is vanishingly small that two mtDNA molecules differing at two or more nucleotide sites would be present in the same mitochondrion or cell and undergo recombination. In addition, cell hybrids containing both human and mouse mtDNA have shown no evidence of recombination, even after several generations (DE FRANCESCO, ATTARDI and CROCE 1980). Thus, although mutation and recombination together serve to generate diversity in nuclear genes, mutation appears to be the primary, and probably sole, source of animal mtDNA sequence variation. These characteristics, together with the relatively recent evolutionary history of the sequences (less than 1% average sequence divergence throughout the molecule; BROWN 1980; FERRIS *et al.* 1981), have, therefore, afforded us a unique opportunity to examine the nature and distribution of evolutionary change at the nucleotide level without the confounding problems resulting from recombination or a large number of multiple substitutions.

In addition, mtDNA sequence variation within and among a variety of organisms has been the focus of several recent studies using restriction endonucleases (e.g., AVISE *et al.* 1979; BROWN, GEORGE and WILSON 1979; SHAH and LANGLEY 1979; BROWN 1980; BROWN and SIMPSON 1981; DENARO *et al.* 1981; LANSMAN *et al.* 1982). The human sequences analyzed in this paper allow us to examine the accuracy of these estimates of sequence variability and divergence.

We have focused our attention on an approximately 900-base pair region which includes the heavy strand origin of replication and the displacement or D-loop (WALBERG and CLAYTON 1981). The region is largely, if not completely, noncoding. A small open reading frame associated with a polyadenylated 7S RNA has been reported in humans (OJALA *et al.* 1981); however, it appears to have been deleted in the mouse (BIBB *et al.* 1981) and cow (ANDERSON *et al.* 1982). We have extended the simple pairwise sequence comparisons of GREENBERG, NEWBOLD and SUGINO (1982) to include a phylogenetic analysis of the sequences and a detailed comparison of estimates of sequence divergence derived from restriction endonuclease studies and direct sequence comparisons. We have also examined in detail evidence for substitutional pathway biases that play a role in the evolution of mtDNA sequences.

MATERIALS AND METHODS

mtDNA sequences: The sequences and alignments are presented in Table 1. Five of the sequences were determined by GREENBERG, NEWBOLD and SUGINO (1982). Two others were obtained from the literature (ANDERSON *et al.* 1981; WALBERG and CLAYTON 1981). Our analysis focuses on an approximately 900-base pair region (position 16129 to 459) for which data are available for all seven sequences. We have used the nucleotide position coordinates for the human mtDNA sequence of ANDERSON *et al.* (1981). The choice of alignments for additions and deletions was unambiguous in all but one case (position 16222). Our decision for this latter instance was based on the overall high degree of similarity between sequences 1 and 7, although other alignments did not significantly alter our conclusions.

Restriction endonuclease analysis: The five mtDNAs sequenced by GREENBERG, NEWBOLD and SUGINO (1982) were isolated from fresh placentas as described. Samples were cleaved with six hexanucleotide-recognizing restriction endonucleases (*EcoRI*, *HpaI*, *HindIII*, *KpnI*, *PstI* and *XbaI*) and fragments separated by electrophoresis as described by GREENBERG, NEWBOLD and SUGINO (1982). Cleavage sites of additional enzymes were obtained from GREENBERG, NEWBOLD and SUGINO (1982) or by examination of the nucleotide sequences. Cleavage sites for another placental mtDNA (DROUIN 1980; ANDERSON *et al.* 1981) and KB cell mtDNA (WALBERG and CLAYTON 1981) were obtained from the source papers and by inspection of the sequences.

Data analysis: We analyzed nucleotide substitutions and sequence length variations separately. Analysis of sequence divergence was carried out in two fundamentally distinct ways. The first method involved simple pairwise comparisons of the proportion of nucleotide positions that differ. This approach does not allow specific identification of convergent substitutions (e.g., $G \rightarrow A \rightarrow G$), parallel substitutions (e.g., $G \leftarrow A \rightarrow G$) or of multiple (but undetected) substitutions at the same position. Various methods have been proposed that attempt to take the occurrence of these events into account in the analysis of pairwise data. We have used the approach of KIMURA (1981; formula 6) since it incorporates differences in the rates of transitions vs. transversions. Thus, the total number of substitutions per nucleotide (K) which have occurred between two sequences since their divergence from a common ancestral sequence is estimated as follows:

$$K = -(1/4) \ln[(1-2P-2Q)(1-2P-2R)(1-2Q-2R)]$$

where P is the relative frequency of homologous sites showing transition base substitutions ($A \leftrightarrow G$ and $C \leftrightarrow T$), and Q and R are the frequencies of sites showing transversions of the type $T \leftrightarrow A$ and $C \leftrightarrow G$ or $T \leftrightarrow G$ and $C \leftrightarrow A$, respectively. The variance of the estimate of K is calculated by equation 12 of KIMURA (1981).

We also analyzed the levels and patterns of differentiation among sequences by a qualitative phylogenetic analysis. For the 899 nucleotides that can be aligned among the seven mtDNA sequences, ten positions contained phylogenetically informative substitutions (i.e., nucleotide differences that are shared by two or more of the sequences; discordancies of FITCH 1977). The distribution of these nucleotides provided the information for the branching order of a phylogenetic network.

TABLE 1
Alignment of human mitochondrial DNA sequences (heavy strand)

Sequence	Position		Sequence
456			G T C
444			V C
316			T
315.1			[G]
302.2			[G]
302.1			[*]
263			T C T
247			C T G
236			C T G
200			T A T
195			T A T
189			T A T
186			T G C
185			T G C
182			T G C
152			T G C
151			T G C
150			T G C
146			T G C
73			T C T
9			T C T
7			T C T
16519			T A T
16424			T A T
16362			T A T
16360			T A T
16356			T A T
16320			T A T
16311			T A T
16304			T A T
16294			T A T
16293			T A T
16280			T T T
16278			T T T
16243			T A T
16242			T A T
16230			T A T
16224			T A T
16223			T A T
16222			T A T
16189			T A T
16188			T A T
16187			T A T
16172			T A T
16167			T A T
16166			T A T
16148			T G C
16134			T G C
16129			T G C

Only positions showing length variation (indicated by brackets) or nucleotide substitutions are shown. Asterisks indicate the absence of a nucleotide in that position. Dashes indicate that the nucleotide is the same as in sequence 1. The region compared is nucleotide 459 to 16129 (through the heavy strand origin of replication at position 191; ANDERSON *et al.* 1981). The numbering system of ANDERSON *et al.* (1981) has been adopted with the exception of 302.1, 302.2 and 315.1. These notations indicate additional nucleotide positions between 302 and 303 and 315 and 316 of the ANDERSON *et al.* sequence (sequence 1), respectively. Arrows identify phylogenetically informative positions (see text). Sequences 2-6 are from clones pBHK2, pDCK1, pCDK1, pLKK3 and pCJ5 of GREENBERG, NEWBOLD and SUCINO (1982), respectively. Sequence 7 is from KB cells (WALBERG and CLAYTON 1981).

Our approach to network construction was basically that outlined by FITCH (1977), and the reader is referred to that paper for a detailed description of the set of operations and their justification. Starting with the phylogenetically informative substitutions, the sequences were linked successively to those that differed by the fewest number of substitutions. Unique substitutions (singularities of FITCH 1977) contribute only to the final branch lengths of the most parsimonious network, i.e., the network requiring the fewest substitutions to relate all sequences. An estimate of the minimum number of substitutions between two sequences can be obtained by counting nucleotide changes proposed along the branches of the network. Details of the analysis and interpretation are presented in Figure 1 and the associated text in RESULTS and DISCUSSION.

Estimates of nucleotide sequence divergence were calculated from fragment patterns according to NEI and LI (1979) by solving the following equality for δ :

$$2n_{xy}/(n_x + n_y) = (e^{-\delta})^4 / (3 - 2e^{-\delta})$$

where n_x and n_y are the total numbers of fragments in sequences X and Y, respectively, and n_{xy} is the number of fragments common to both sequences. Sequence divergence was estimated from restriction site maps by the approach of KAPLAN and RISK0 (1981) by solving the following equation for $2\bar{\eta}$ equivalent to δ of NEI and LI (1979):

$$\sum_{k_i=4}^6 \frac{\bar{X}_1(k_i)}{2} e^{-k_i 2\bar{\eta}} = \sum_{k_i=4}^6 \bar{X}_2(k_i)$$

where k_i is the number of nucleotides in the restriction enzyme recognition site (here we have used enzymes with sites of four, five and six nucleotides), and where $\bar{X}_1(k_i)$ is the average number of

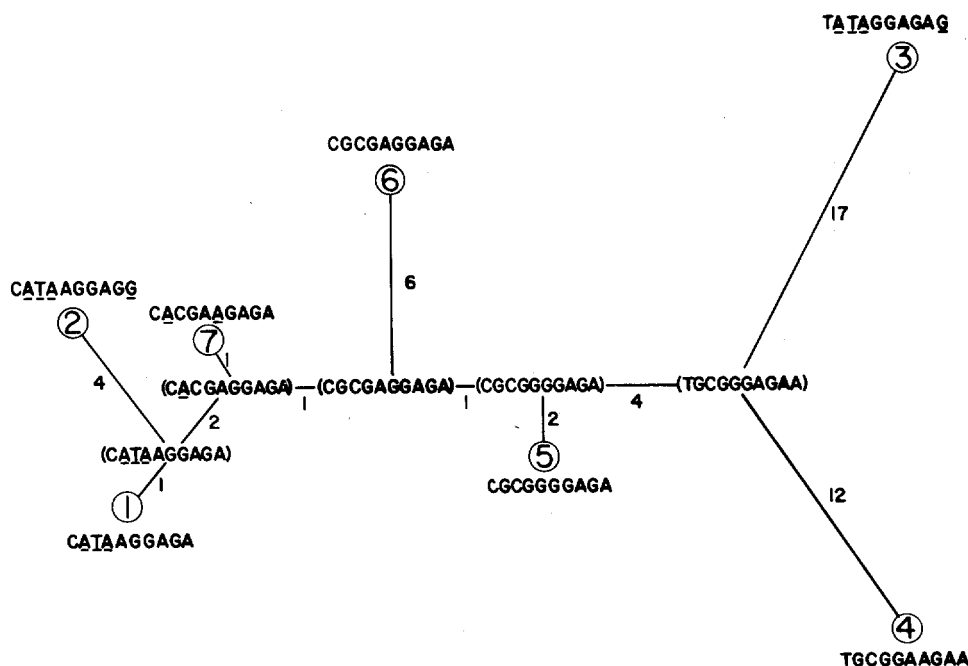


FIGURE 1.—A phylogenetic network constructed from the seven human mtDNA sequences according to maximum parsimony criteria. Only the ten phylogenetically informative nucleotide positions (see Table 1) are shown. Sequences in parentheses represent hypothetical intermediates. Numbers along branches are the number of substitutions required along the legs of the network. Underlining denotes convergent or parallel substitutions.

cleavage sites k_i nucleotides long, and $\bar{X}_2(k_i)$ is the number of cleavage sites of length k_i shared by the sequences being compared. This formulation reduces to equation 8 of NEI and LI (1979) when all restriction enzymes recognize sites of identical length but allows a single estimate of divergence to be made when enzymes of different recognition site lengths are used. The variance of $2\hat{\eta}$ is estimated as 4 times the variance of $\hat{\eta}$ (KAPLAN and RISK0 1981; equivalent to variance of δ given by NEI and TAJIMA 1981).

RESULTS

Pairwise nucleotide sequence comparisons: Alignment of the human mtDNA sequences reveals sequence length variation at three locations (Table 1). The putative additions or deletions are each small in size (one or two nucleotides) and are generally located within multimeric repeats (details in GREENBERG, NEWBOLD and SUGINO 1982). Nucleotide sequence differences among the human mtDNAs are summarized in Table 2. These values are based on comparisons of 899 nucleotides from each sequence; positions involved in additions or deletions were not included. Between four and 28 nucleotides differed among the sequences within this region. This corresponds to estimates of the adjusted total number of base substitutions per nucleotide (K ; KIMURA 1981) ranging from 0.004 to 0.032 (mean of 0.017; Table 2).

Phylogenetic network estimates of divergence: A parsimonious phylogenetic network was constructed to provide additional estimates of the number of substitutions. An important advantage of this type of analysis is the prediction of probable intermediate sequences. These hypothesized sequences reveal convergent and parallel substitutions that are not detected by making simple pairwise comparisons of extant sequences.

The most parsimonious network is easily obtained among sequences 1, 2, 5, 6 and 7 since they contain no parallel or convergent substitutions (discordancy diagram analysis among these sequences reveals no unavoidable discordancies or homoplasy; see FITCH 1977 for a description of this analysis). The number of substitutions required by this network is equal to the minimum number of pairwise differences between these five sequences. The addition of sequences 3 and 4 to this network is more difficult, however, since the addition of either sequence alone, as well as both together, results in parallelism or convergence. Indeed, discordancy diagram analyses reveal that among all seven sequences there is a minimum of five unavoidable discordancies, meaning that the number of substitutions along the branches of the most parsimonious network among the seven sequences will be equal to the minimum number of differences between the sequences (46) plus five parallel and/or convergent substitutions for a total of 51 substitutions. Since we know the most parsimonious network for five of the seven sequences, it is not difficult to enumerate all branching combinations of sequences 3 and 4 with this network in order to determine the structure of the most parsimonious network among all seven sequences. Only one network, shown in Figure 1, was found to contain 51 substitutions; all others required 52 or more base changes.

Since many different branching orders for the seven sequences differed from the most parsimonious network by only one or a few additional substitutions, it would be unwise to ascribe very great significance to this most parsimonious tree as necessarily reflecting the true phylogenetic relationships among the

TABLE 2

Observed number of nucleotide differences and estimates of the total number of base substitutions among human mitochondrial DNA sequences in the 899-nucleotide region of comparison

Sequence	1	2	3	4	5	6	7
A.							
1	—	5	20	21	7	10	4
2	0.006 (0.003)	—	21	24	10	13	7
3	0.023 (0.005)	0.024 (0.005)	—	28	23	28	22
4	0.024 (0.005)	0.027 (0.006)	0.032 (0.006)	—	18	23	17
5	0.008 (0.003)	0.011 (0.004)	0.026 (0.005)	0.020 (0.005)	—	9	5
6	0.011 (0.004)	0.015 (0.004)	0.032 (0.006)	0.026 (0.005)	0.010 (0.003)	—	8
7	0.004 (0.002)	0.008 (0.003)	0.025 (0.005)	0.019 (0.005)	0.006 (0.003)	0.009 (0.003)	—
B.							
1	—	5	26	21	7	10	4
2	5.0	—	29	24	10	13	7
3	20.4	21.5	—	28	23	28	24
4	21.5	24.6	28.9	—	18	23	19
5	7.0	10.1	23.6	18.3	—	9	5
6	10.1	13.2	28.9	23.6	9.1	—	8
7	4.0	7.0	22.6	17.4	5.0	8.1	—

Matrix A: Observed number of nucleotide differences (above diagonal) and estimates of the number of base substitutions per nucleotide (K) calculated according to KIMURA (1981) (below diagonal).

Matrix B: Estimates reflecting numbers of substitutions along branches of a phylogenetic network (Figure 1) are above the diagonal; below the diagonal are estimates derived from adjusted pairwise comparisons using KIMURA's (1981) approach.

seven human mtDNAs. This situation was due, in large part, to a predominance of transitions resulting in the occurrence of parallel and convergent substitutions at half of the phylogenetically informative nucleotide positions. However, the validity of the conclusion concerning the minimum number of parallel and/or convergent substitutions stands firmly since a minimum of five must be invoked among the ten phylogenetically informative substitution positions regardless of the network chosen. This conclusion is further confirmed by the discordancy diagram analyses which are independent of network construction.

Thus, convergent or parallel base changes occur at a minimum of five of the 45 nucleotide substitution sites (11%). An additional multiple base substitution is apparent at a sixth site, but since three different nucleotides are found at the same position in the different sequences, it is classified as a divergent multiple substitution.

The minimum number of substitutions along the branches between pairs of sequences are enumerated in Table 2 and compared with estimates based on

simple pairwise comparisons of the sequences. The branch lengths represent *minimal* estimates of both divergent and convergent/parallel substitutions since the network is based on the assumption of parsimony. Additional multiple substitutions may have occurred but remain undetected due to reversion to the same nucleotide in the available sequences. It is clear that even for the low levels of divergence among these sequences, the high proportion of convergent and parallel substitutions can lead to a significant underestimate of the true rates of substitution. For example, a pairwise comparison between sequence 2 and 3 shows only 21 base differences, yet the sequences actually differ by at least 29 substitution events. Methods that attempt to account for multiple substitutions (e.g., KIMURA 1981) predict only 21.5 substitutions on average between these two sequences (Table 2). For the low level of divergence characteristic of intraspecific comparisons in humans, other metrics, including those not taking into account differences in substitution pathway rates (e.g., KAPLAN and RISK0 1982), gave virtually identical estimates.

Substitution pathway biases: The types of substitutions predicted from the phylogenetic analysis are presented in Table 3. If all 12 base pair substitution pathways were equally probable, transversions would be twice as frequent as transitions. However, as observed by GREENBERG, NEWBOLD and SUGINO (1982) in their presentation of the human mtDNA sequences, transitions appear to predominate over transversions. In our study, 49 of 51 substitutions (96.1%) are transitions (including the five convergent and parallel substitutions). To reveal substitution biases that occur in addition to the apparent transition bias, we have broken down and analyzed the observed substitution types as suggested by BROWN and CLEGG (1983).

We first tested the null hypothesis that there is no bias in the susceptibility of a nucleotide having been replaced, *i.e.*, that no nucleotide or nucleotides are more likely to be the site of a substitution than the others (Table 4). The expected values are derived from the base composition of the hypothesized ancestral sequence. None of the comparisons shows a significant departure from the expected. However, purines (particularly G) tend to be replaced more often than pyrimidines (particularly T) on the H strand (C more than A on the L strand).

Similarly, we can test the null hypothesis that given that a base has changed, then no bias exists as to the nucleotide introduced into that site (Table 5). Here the expectations are calculated on the basis of the observed frequency of nucleotides in the ancestral sequence or, alternatively, on the basis of the observed frequency of nucleotides replaced. This second approach incorporates the twofold bias for the purine G over the pyrimidine C as the site of a nucleotide substitution on the H strand. It is clear from both comparisons that a significant bias in the nucleotide introduced on the H strand is indicated, with an approximately 2.5-fold preference shown for purines. There is, of course, a corresponding bias favoring pyrimidines on the complementary L strand. No other biases are evident.

Our remaining concern is the detection of additional biases affecting the frequency of individual substitution pathways. We have treated Table 3 as a 4

TABLE 3

Observed single base differences among human mitochondrial DNA sequences^a

Ancestral sequence		Nucleotide introduced				Total
Nucleotide replaced	No. in sequence	A	G	C	T	
A	214	—	13	0	0	13
G	285	20	—	1	0	21
C	127	0	0	—	7	7
T	272	0	1	9	—	10
Total	899	20	14	10	7	51

^a Obtained from our phylogenetic network for the seven H strand sequences. The best estimate of the ancestral sequence for the seven mtDNAs was taken to be the hypothetical sequence represented by the node between sequences 3 and 4 in Figure 1.

TABLE 4

Observed and expected number of nucleotides replaced from the ancestral H strand sequence

Nucleotide replaced	Observed no.	Expected no. ^a	$\chi^2_3 = 3.43$
A	13	12.14	
G	21	16.17	
C	7	7.19	
T	10	15.45	
Component		d.f.	χ^2
Purine (A + G)	vs. Pyrimidine (C + T)	1	2.55
2 H-pair (A + T)	vs. 3 H-pair (G + C)	1	1.69
A + C	vs. G + T	1	0.04

^a The expected number is calculated as the total number of substitutions observed (51) times the frequency of the respective nucleotide in the ancestral sequence.

× 4 incomplete contingency table (BROWN and CLEGG 1983). Expectations are calculated from the marginal totals and, thus, take into account the biases favoring purines as the nucleotides replaced from and introduced into the ancestral sequence. We must, however, adjust the marginal totals for the unobserved substitutions in the diagonal cells (e.g., A → A, G → G, etc.; see footnote to Table 6). The expected frequencies for the individual substitution pathways (the ij^{th} cells of Table 3) are given in Table 6. Biases favoring purines (or pyrimidines) as both the site of substitution and the nucleotide introduced will lead to a ratio of transitions to transversions greater than one-half, the ratio expected if no biases existed (e.g., the ratio is 0.77 in our data; Table 6). However, in the expectations given in Table 6, these biases have been taken into account. Nonetheless, a highly significant bias is still evident favoring transitions over transversions (by a factor of almost 32 times).

Spatial distribution of substitutions: Two clusters of sites of substitution are apparent, roughly encompassing the ends of the D-loop region (GREENBERG, NEWBOLD and SUGINO 1982), suggesting that the rate of substitution is not

TABLE 5

Observed and expected number of nucleotides introduced into the ancestral H strand sequence

Nucleotide introduced	Observed no.	Expected no.		
		(a)	(b)	
A	20	12.94	12.67	
G	14	11.59	10.00	
C	10	14.59	14.67	
T	7	11.83	13.67	
		χ^2		
Component		d.f.	(a)	(b)
Purine (A + G) vs. Pyrimidine (C + T)		1	7.02**	10.20**
2 H-pair (A + T) vs. 3 H-pair (G + C)		1	0.38	0.03
A + C vs. G + T			0.47	0.56
Total		3	7.77	10.58*

Expected condition on: (a) the observed frequency of nucleotides in the ancestral sequence. For example, the expected number of As is equal to one-third of the substitutions expected from the other three nucleotides taking base composition into account; $51[(285/899)/3] + 51[(127/899)/3] + 51[(272/899)/3] = 12.94$. (b) the observed frequency of nucleotides replaced. For example, the expected number of As is equal to one-third of the observed substitutions at the other three nucleotide sites; $(21 + 7 + 10)/3 = 12.67$.

* $P < 0.05$
** $P < 0.01$

uniform for all sites. Under the null hypothesis that the sites of substitution are randomly distributed, the distribution of runs of consecutive invariant sites should be geometric (BROWN and CLEGG 1983) with mean q/p and variance q/p^2 , where p is the probability of a site having a substitution in at least one sequence and q is the probability that a site is identical in all sequences. For our data, $p = 45/899 = 0.05$ and $q = 1 - p = 0.95$. The observed variance in run length is 799.26, significantly larger than expected (379.07; $\chi^2_{43} = 90.67$, $P < 0.001$, p. 117 in BREIMAN 1973). Calculation of the expected frequency for each run of length r (npq^r , where $n = 45$ is the total number of altered sites) reveals that the large observed variance in run length is due to an excess of both adjacent sites of substitution ($r = 0$, apparent "hot spots") and long runs of consecutive invariant sites (apparent conserved blocks; analysis not shown).

We can also test the hypothesis that the distribution of the number of substitutions per site is consistent with that of a uniform rate of substitution. Following FITCH and MARKOWITZ (1970; considering nucleotide substitutions rather than amino acid replacements, however) we assume that all nucleotide sites are equally variable. The expected distribution for the number of substitutions per site then becomes the Poisson. Comparison with the observed distribution reveals a significant excess of multiple substitutions (Table 7). Thus, we can conclude that the rate of substitution is significantly nonuniform in the D-loop region of human mtDNA, resulting in a clustering of variable sites and an excess of multiple (repeated) substitutions.

Restriction site and fragment comparisons: Recognition sites along the entire

TABLE 6

Analysis of nucleotide substitutions

Type of substitution	Change in H bonding	No. observed	No. expected ^a
Transitions			
A → G } T → C }	2 → 3 } ΔH ≠ 0	13 } 22 } 9 } 49	7.29 } 9.09 } 1.80 } 22.47
C → T } G → A }	3 → 2 }	7 } 27 } 20 }	0.98 } 13.38 } 12.40 }
Transversions			
A → T } T → A }	2 → 2 } ΔH = 0	0 } 0 } 0 } 1	2.51 } 7.01 } 4.50 } 14.80
C → G } G → C }	3 → 3 }	0 } 1 } 1 }	2.84 } 7.79 } 4.95 }
A → C } T → G }	2 → 3 } ΔH ≠ 0	0 } 1 } 1 } 1	3.29 } 7.27 } 3.98 } 14.26
C → A } G → T }	3 → 2 }	0 } 0 } 0 }	3.21 } 6.99 } 3.78 }
Component		d.f.	χ ²
Between all pathway classes		5	100.48***
Between 6 classes		4	57.53***
3 types of substitution		2	56.52***
Transitions vs. transversions		1	56.52***

The expected number of substitutions is based on the marginal frequencies of Table 1 and, thus, takes account of biases in the nucleotide replaced from and introduced into the ancestral H strand sequence (see text).

^a The expected frequency for the ij^{th} cell in Table 3 is $p_i(m_i + f_i)$, where i and j refer to row and column number, respectively, m_i is the observed number of times the i^{th} nucleotide was replaced, f_i is the unobserved diagonal cell of the i^{th} row, and p_i is the probability that the newly introduced base will be an A, G, C or T ($j = 1, 2, 3$ or 4 , respectively). The value of f_i is calculated as $kp_i - n_i$, where k is the total number of substitutions, including both observed (51) and unobserved (diagonal cell) changes, and n_i is the observed number of times the i^{th} nucleotide was introduced into the ancestral sequence. Values for k and p_i are obtained by interactively solving the quadratic $kp_i^2 - (k + n_i - m_i)p_i + n_i = 0$ such that $0 \leq p_i \leq 1$ and $\sum p_i = 1$ (Brown and CLEGG 1983). For our data, $k = 72.78$ and the $\{p_i\}$ are 0.3881, 0.3430, 0.1550 and 0.1182.

* $P < 0.001$

mtDNA for six hexanucleotide-recognizing restriction endonucleases are presented in Table 8. Restriction site data were not available for the entire KB cell mtDNA (sequence 7). The only variation observed was due to the presence or absence of two different KpnI cleavage sites. The mean number of substitutions per nucleotide estimated from the maps ranged from 0 to 0.010 with a mean of 0.006 (Table 9). Estimates from fragment pattern comparisons would be approximately similar provided all fragment size variation could be resolved. However, due to the large size of the KpnI A fragment (13.6 kb), an 80-base variation was not resolved, which biased estimates downward (mean of 0.002; Table 9) prior to our accession of nucleotide sequence data.

All commercially available restriction endonucleases with recognition sites in

TABLE 7
Distribution of the number of substitutions per nucleotide

	No. of substitutions					χ^2	d.f.	P
	0	1	2	3	4			
Observed	854	39	6	0	0			
Expected ^a	849.42	48.18	1.37	0.03	0.00	1059.5 ^b	898	0.001

^a Calculated as $ne^{-\bar{i}}\bar{i}^i/i!$ where $n = 899$, the total number of sites; $\bar{i} = 51/899$, the mean number of substitutions per site; and i is the number of substitutions at a particular site.
^b $\chi^2 = (1/i) \sum_{i=0}^4 x_i - \sum_{i=0}^4 ix_i$, with d.f. = $899 - 1$, where x_i is the observed number of sites with exactly i substitutions (FITCH and MARKOWITZ 1970).

TABLE 8
Location and distribution among human mtDNA molecules of recognition sites for six hexanucleotide-recognizing restriction endonucleases

Restriction sites	mtDNA sequence						Convergent or parallel site gain or loss?
	1	2	3	4	5	6	
Variable sites							
KpnI (16129-34)	+	+	-	+	+	-	Loss
KpnI (16048-53)	+	-	+	-	-	+	No
Invariant sites							
EcoRI (4121-6; 5274-9; 12640-5)							
HpaI (5691-6; 10014-9; 12406-11)							
Hind III (6203-8; 11680-5; 12570-5)							
KpnI (2573-8)							
PstI (6910-5; 9020-5)							
XbaI (1193-8; 2953-8; 7440-5; 8286-91; 10256-61)							

Numbers in parentheses are the location of recognition sequences on the human mtDNA sequence of ANDERSON *et al.* (1981). Restriction enzyme cleavage of mtDNAs 2-6 produced fragments consistent in size with those predicted by sites at the same locations in the ANDERSON *et al.* sequence (no. 1) and reported by DROUIN (1980).

the 899-base pair region are listed in Table 10, together with the distribution of the sites among the sequences. The comparison of all sites, without regard to the potential ability to resolve those that are closely positioned, yields estimates of sequence divergence ranging from 0.006 to 0.032, with a mean 0.019 (Table 11). Under the more realistic assumption that sites less than 50 nucleotides apart cannot in practice be resolved, the restriction site estimates of divergence drop to a mean of 0.012 (range of 0.004-0.022; Table 11).

DISCUSSION

Nucleotide sequence divergence in human mitochondrial DNA: BROWN (1980) and FERRIS *et al.* (1981) have estimated the average proportion of nucleotide differences in the human populations to be 0.003-0.004 for the whole mtDNA molecule based on restriction endonuclease fragment patterns. Our restriction site and fragment analysis of human mtDNA molecules with six hexanucleotide-recognizing endonucleases yielded similar estimates. However, direct compari-

TABLE 9

Number of nucleotide substitutions per base among human mtDNA molecules estimated from restriction maps and fragment pattern data for six hexanucleotide endonucleases

Sequence	1	2	3	4	5	6
1	—	0.005 (0.005)	0.005 (0.005)	0.005 (0.005)	0.005 (0.005)	0.005 (0.005)
2	0.005 0.002	—	0.010 (0.007)	0.000	0.000	0.010 (0.007)
3	0.005 0.005	0.007 0.003	—	0.010 (0.007)	0.010 (0.007)	0.000
4	0.005 0.002	0.000 0.000	0.007 0.003	—	0.000	0.010 (0.007)
5	0.005 0.002	0.000 0.000	0.007 0.003	0.000	—	0.010 (0.007)
6	0.005 0.005	0.007 0.003	0.000 0.000	0.007 0.003	0.007 0.003	—

Estimates from restriction maps (and associated standard errors in parentheses) are above the diagonal; those below are from fragment pattern comparisons. The upper fragment estimates assume all size differences are resolved; whereas the lower fragment estimates assume differences in the large *KpnI* fragment could not be detected (see text). The statistical properties of fragment estimates are not known, although the associated sampling error is probably larger than that for restriction site map estimates (NEI and LI 1979).

sons of sequences examined in the present paper provide estimates for an 899-base pair region (5.4% of the mtDNA genome) that are several fold higher (0.017, with a range of 0.004–0.032). This disparity is consistent with the rapidly evolving nature of this region (UPHOLT and DAWID 1977; WALBERG and CLAYTON 1981; GREENBERG, NEWBOLD and SUGINO 1982).

Although this apparent agreement between actual sequence comparisons and estimates based on restriction site variability for human mtDNA is encouraging, it conceals some very important biases. These shortcomings of restriction surveys become apparent upon comparison of the nucleotide sequences and the actual substitutions responsible for the loss or gain of restriction sites.

Restriction site variability: The average sequence divergence among several closely related mtDNAs does appear to be reasonably estimated by restriction analyses. However, restriction site surveys may not resolve regions showing differing levels of variability. The two hypervariable domains in the D-loop region have not produced a detectable clustering of variable restriction sites in the D-loop region of primates (BROWN and GOODMAN 1979; FERRIS, WILSON and BROWN 1981; FERRIS *et al.* 1981) or rodents (BROWN *et al.* 1981; LANSMAN *et al.* 1982). This probably results, at least in part, from the nonrandom distribution of cleavage sites along the mtDNA molecule (ADAMS and ROTHMAN 1982).

It is also clear that at the low levels of divergence typical of human mtDNA, estimates of divergence based on restriction data may poorly reflect actual sequence similarity, particularly when only a small number of endonucleases are used. Although our region of direct sequence comparison is only 5.4% of the total mtDNA genome and appears to be faster evolving than the remainder of the molecule (UPHOLT and DAWID 1977; GREENBERG, NEWBOLD and SUGINO 1982),

TABLE 10

Location and distribution within the 900-base pair mtDNA region of recognition sites for commercially available restriction endonucleases

Variable sites	mtDNA sequence							Convergent or parallel site gain or loss?
	1	2	3	4	5	6	7	
4-base sites								
<i>Fnu4HI</i> (260-4)	—	+	+	+	+	+	+	No
<i>MboI</i> (8-11)	—	—	+	—	—	—	—	No
<i>HaeIII</i> (16517-20)	—	—	—	+	+	+	+	No
<i>Sau96I</i> (16516-20)	—	—	—	+	+	+	+	No
<i>RsaI</i> (16303-6)	+	—	+	+	+	+	+	No
<i>RsaI</i> (16310-3)	+	+	—	—	—	+	+	No
<i>MnII</i> (144-7)	+	+	+	+	—	+	+	No
<i>MnII</i> (16355-8)	—	—	—	—	—	+	—	No
<i>MnII</i> (16222-5)	+	+	—	—	—	+	+	Loss
<i>MnII</i> (16187-90)	+	+	—	—	+	+	+	Loss
<i>MnII</i> (16185-8)	—	—	—	+	—	—	—	Gain
<i>TaqI</i> (16278-81)	—	—	—	+	—	—	—	No
5-base sites								
<i>HphI</i> (184-8)	—	—	—	+	—	—	—	No
6-base sites								
<i>BstEII</i> (184-90)	—	—	—	+	—	—	—	No
<i>KpnI</i> (16129-34)	+	+	—	+	+	—	+	Loss
Invariant sites								
4-base sites								
<i>AluI</i> (37-40; 16476-9)	<i>MnII</i> (308-11; 16446-9; 16407-10; 16379-82; 16261-4)							
<i>DdeI</i> (16380-4)	<i>RsaI</i> (16334-7; 16208-11; 16156-9; 16130-3)							
<i>HaeIII</i> (322-5; 16456-9)	<i>Sau3A</i> and <i>MboI</i> (1-4)							
<i>HinfI</i> (136-40)	<i>Sau96I</i> (16456-60; 16455-9; 16390-4)							
<i>HpaII</i> (104-7; 16453-6)	<i>ThaI</i> (78-81)							
5-base sites								
<i>AvaII</i> (16390-4)	<i>HphI</i> (430-4; 14-18; 16263-7)							
<i>HgaI</i> (94-98)								
6-base sites								
<i>BalI</i> (321-6)	<i>SstI</i> (36-41)							
<i>HgiAI</i> (107-12)								

Numbers in parentheses are the location of the recognition sequence on the heavy strand sequence. Presence or absence of a site in a sequence is indicated by a plus (+) or minus (-), respectively.

there is no reason to assume that divergence in this segment should not reflect the relative ranking among the sequences as to average mtDNA divergence. However, virtually no correlation exists between observed sequence differences and estimates based on either whole mtDNA restriction site maps or on comparisons of fragment patterns for six hexanucleotide-recognizing endonucleases (Table 9). A case in point involves sequences 3 and 6, which differ at 3.1% of their nucleotides; yet, the mtDNAs from which they were cloned share identical restriction maps, including a shared loss of a *KpnI* site relative to the other sequences. Examination of the base substitutions responsible for this shared site loss revealed two distinct substitutions at different positions within

TABLE 11

Estimates of sequence divergence per nucleotide (and associated standard errors)
in the 899-base pair region computed from restriction site data in Table 10

Sequence	1	2	3	4	5	6	7
1	—	0.006 (0.004)	0.021 (0.009)	0.032 (0.011)	0.020 (0.008)	0.016 (0.007)	0.009 (0.005)
2	0.004 (0.004)	—	0.021 (0.009)	0.032 (0.011)	0.020 (0.008)	0.016 (0.007)	0.009 (0.005)
3	0.012 (0.007)	0.008 (0.006)	—	0.026 (0.010)	0.021 (0.009)	0.023 (0.009)	0.023 (0.009)
4	0.022 (0.009)	0.018 (0.008)	0.015 (0.007)	—	0.019 (0.008)	0.028 (0.010)	0.021 (0.008)
5	0.016 (0.008)	0.011 (0.007)	0.019 (0.009)	0.014 (0.007)	—	0.016 (0.007)	0.009 (0.005)
6	0.016 (0.008)	0.011 (0.007)	0.012 (0.007)	0.014 (0.007)	0.007 (0.005)	—	0.006 (0.004)
7	0.011 (0.007)	0.007 (0.005)	0.015 (0.008)	0.010 (0.006)	0.004 (0.004)	0.004 (0.004)	—

Restriction estimates were calculated using all available endonuclease sites, identified in Table 10, without regard to the potential ability to resolve closely positioned sites (above diagonal), and under the more realistic assumption that only sites separated by at least 50 nucleotides could be resolved, on average, by conventional restriction mapping procedures (below diagonal).

the *KpnI* recognition site between these two sequences (position 16129 in sequence 3, and position 16134 in sequence 6). Thus, these sequences actually differ at this recognition site by at least two substitutions.

The use of all possible restriction sites in the 900-base pair region, without regard to the potential ability to resolve closely positioned sites (approximately 37 sites per sequence; Table 11), yields estimates reasonably well correlated with sequence divergence (Figure 2). In practice, however, restriction analysis estimates may be biased downward due to the inability to distinguish small differences in restriction site position or fragment size. For example, BROWN, GEORGE and WILSON (1979) noted that the average limit of site resolution in their comparisons of primate mtDNAs was approximately 1% of the genome (165 nucleotides). If one assumes that sites less than 50 nucleotides apart could not be distinguished, sequence divergence in the 900-base pair region of human mtDNA is progressively underestimated as divergence increases (Figure 2). Three of the 21 restriction estimates are more than two standard errors smaller than the observed degree of sequence divergence.

The occurrence of convergent or parallel substitutions at a single nucleotide within a recognition sequence (e.g., site of *HaeIII* and *Sau96I* at 16516-20) also contributes to an underestimation of sequence divergence in restriction analyses due to a flickering on and off of these restriction sites in various evolutionary lineages. This apparent "flickering" has been observed at several restriction sites in an extensive study of variation in deer mice mtDNA (LANSMAN *et al.* 1982). We have also observed several examples of convergent site loss (*KpnI* at 16129-34 and two of the five variable *MnII* sites; Table 10) and a case of adjacent substitutions destroying one *MnII* site and creating a new site two nucleotides

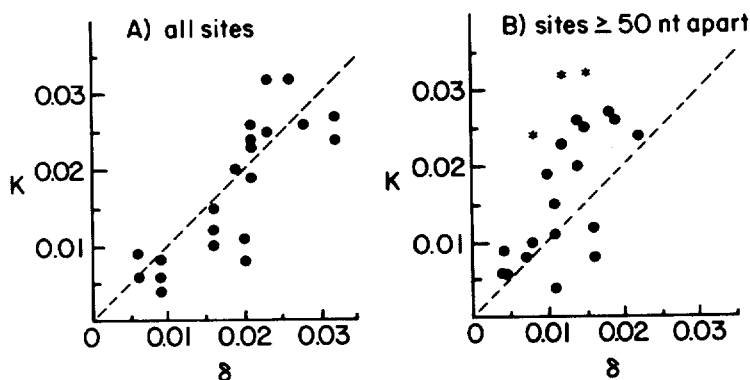


FIGURE 2.—Comparison of sequence divergence in the 900-base pair region estimated from nucleotide sequence comparisons (K) and from restriction site maps for all commercially available restriction endonucleases (δ) assuming (A) all site differences are resolved, and (B) only sites separated by 50 or more nucleotides could be distinguished. The dashed line represents the expectation if both methods revealed identical patterns and levels of differentiation. Comparisons in which the sequence estimate was more than two standard errors from the restriction estimate are indicated by asterisks, whereas those less than two standard errors are indicated by solid dots.

away (at 16185-8). This difference would have been indistinguishable without direct sequence data.

Most methods for estimating the number of nucleotide substitutions from restriction fragment and map data incorporate the probability of multiple mutational events within an endonuclease recognition site under the assumption that both substitutions and restriction sites are randomly distributed along the sequence. The violation of these assumptions for mtDNA suggests that these estimation methods may not account sufficiently for the number of substitutions within a site. Recognizing this potential bias, NEI and LI (1979; equation 15) proposed an estimation procedure that incorporates varying rates of substitution between restriction sites. However, the magnitude of the variance in substitution rates that should be used is generally not known. Additional biases in the estimation of sequence divergence may result from unequal nucleotide frequencies and substitution pathway biases (e.g., the high bias for transitions in mtDNA). The effect of this on pairwise estimates is relatively small at low levels of divergence (TAJIMA and NEI 1982; KAPLAN 1983). However, the occurrence of parallel and convergent restriction site gain and loss at a relatively high rate will be a more serious limitation in the construction of qualitative phylogenies derived from the maps themselves (for an example, see LANSMAN *et al.* 1982).

Estimation from restriction data of the proportion of nucleotide sites that are polymorphic (\hat{p}) in a sample of sequences (ENGELS 1981; EWENS, SPIELMAN and HARRIS 1981; HUDSON 1982) will be subject to similar practical limitations and biases. However, these methods estimate the proportion of nucleotide sites that differ in one or more of the sequences, rather than the total number of substitutions that have occurred since the divergence of the sequences. Thus, a high frequency of parallel and convergent substitutions will probably have little effect on the accuracy of the estimates. Estimates of θ (the product of the

effective population size and mutation rate) derived from \hat{p} will, however, be biased because the actual substitution rate will have been underestimated. (The infinite-allele assumption underlying the estimation of θ from \hat{p} is violated by the occurrence of parallel and convergent substitutions.) At the low levels of divergence typical of intraspecific comparisons, this bias will likely be small. Hence, the actual value of θ will probably not be significantly different than that estimated.

We have not raised the aforementioned points merely as a criticism of previous restriction analyses or as an admonition against the use of restriction analyses to assay sequence variation. These observations do, however, emphasize the importance of recognizing that the standard errors of restriction estimates are very large (often on the order of the estimates; Tables 9 and 11), and that appropriate caution must be used when interpreting phylogenies for closely related sequences based on restriction site variation, particularly when only a small number of sites is examined. Our comparisons also suggest that, in practice, restriction site and fragment comparisons will more likely lead to an underestimate, rather than an overestimate, of actual nucleotide sequence divergence (although see BROWN and SIMPSON 1982). It is of interest to note that all restriction site variation observed in this study was due to nucleotide substitution; no additions or deletions occurred within endonuclease recognition sequences.

Estimation of divergence from sequence data: Our phylogenetic analysis revealed that current methods for using pairwise comparisons to estimate the total number of substitutions that have occurred since two sequences diverged from a common ancestral sequence often fail to account adequately for the actual number of multiple substitutions. These methods are based on models derived using several assumptions that are violated to some degree in the region of mtDNA we have analyzed. For example, base composition is not uniform, the H strand base composition in this region being 24% A, 32% G, 14% C and 30% T. More importantly, there are several significant biases in the frequencies of substitution pathways, most notable being the strong bias favoring transitions over transversions. Several workers have attempted to incorporate these characteristics into estimation procedures (for example, KIMURA 1981; BROWN *et al.* 1982; GOJOBORI, ISHII and NEI 1982). However, unless divergence is large, the effect of these biases alone on the estimation of sequence divergence is relatively small (GOJOBORI, ISHII and NEI 1982; KAPLAN 1983). This is also demonstrated by the fact that estimates of divergence in the human mtDNA D-loop region are virtually identical whether or not the transition bias is included. As with restriction map and fragment analyses, inequality of rates of substitution among nucleotide positions in the sequence appears to be the assumption whose violation has led to the largest bias in divergence estimation.

Length variation and phylogeny: The sequence length variation observed among the humans analyzed here generally occur in repeated mononucleotide or dinucleotide stretches of sequence (GREENBERG, NEWBOLD and SUGINO 1982). This repeated structure of nucleotide sequence associated with additions and deletions is consistent with a model of frameshift mutagenesis proposed by

STREISINGER *et al.* (1966). It follows from this model that these regions of the genome may be highly susceptible to repeated insertion and deletion events (e.g., OKADA *et al.* 1972). Consequently, the distribution of sequence length variation among sequences may not accurately reflect the phylogenetic relationships of the sequences examined. We have tested this hypothesis by superimposing the distribution of additions and deletions on the phylogeny of nucleotide substitutions (Figure 3). The relative ranking of sequence similarity based on substitutions vs. length variation are not in close accordance. For example, sequences 1 and 7 differ by fewer base substitutions than any other two sequences, yet show at least three separate additions or deletions. In contrast, two of the most divergent sequences in terms of nucleotide substitution (3 and 6) show no length differences at all. Part of this length similarity appears to be due to the parallel addition of a single base pair between nucleotides 302 and 303 in the separate lineages leading to sequences 3 and 6 (Figure 3). Other putative sequences at the internal branch nodes all lead to at least one instance of convergence or parallelism for addition/deletion events. Restructuring the tree to avoid these convergent or parallel events yields arrangements that require additional convergent and parallel base substitutions and are, therefore, less parsimonious. It is not clear how insertions and deletions should be weighted relative to nucleotide substitutions in the construction and choice of the most parsimonious phylogenetic network. However, given the mechanism apparently generating sequence length variation in human mtDNA, the shared presence of small additions or deletions may not accurately reflect propinquity of descent due to a potentially high rate of convergence and parallelism. Nonetheless, the distributions of substitutions and addition/deletion events among human mtDNA sequences are reasonably concordant when viewed phylogenetically as in Figure 3.

Mechanisms and rates of mitochondrial DNA evolution: Several workers have suggested that mammalian mtDNA evolution occurs primarily by base substitution, with sequence rearrangements being limited to relatively rare additions and/or deletions (e.g., BROWN 1981; LANSMAN *et al.* 1982). These conclusions are based largely on restriction site and fragment variation studies, with only limited data from direct sequences. Our intraspecific human sequence comparison suggests that small additions and deletions comprise a small, although significant proportion of the total sequence alterations in the D-loop region (five of 56 or 8.9%). However, the high density of coding sequence outside the D-loop region (ANDERSON *et al.* 1981, 1982; BIBB *et al.* 1981) suggests that these length variations must be largely confined to the noncoding region we examined or to the short intergenic stretches. In addition, the insertions and deletions that have been observed among more divergent mammals occur primarily in these noncoding regions (ANDERSON *et al.* 1981, 1982; BIBB *et al.* 1981; FERRIS, WILSON and BROWN 1981; WALBERG and CLAYTON 1981).

Our comparisons of recently diverged sequences have also revealed some important new insights into the nature of the substitution process. The types of substitutions observed are significantly nonrandom. Most notable is an extraordinarily high bias (32-fold) for transitional substitutions in human mtDNA. In

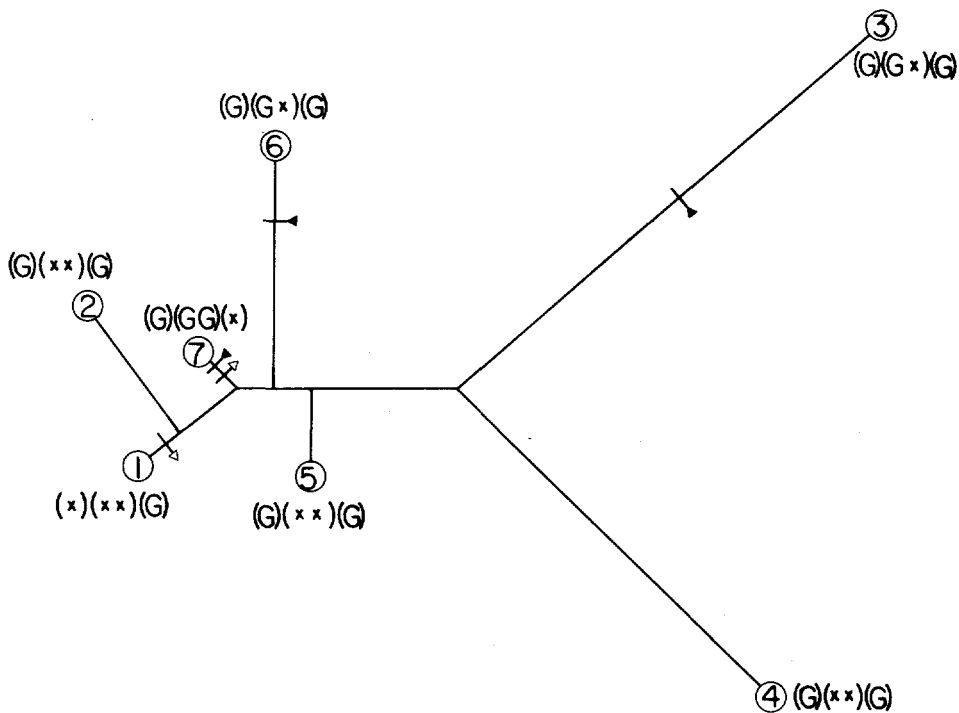


FIGURE 3.—The distribution of sequence length variation among the human mitochondrial DNA sequences together with the parsimonious phylogeny derived from base substitution data. The three locations of length variation shown in parentheses are those indicated in Table 1. The internal branch points of the tree are all hypothesized to have had the following sequence: (G) (**) (G). The short intersecting lines denote hypothesized deletions (Δ) or additions (\blacktriangle).

addition, guanine (a purine) appears to be slightly favored as the site of substitution on the H strand (C on the L strand). Purines are also significantly favored as the nucleotide introduced at the site of substitution on the H strand (pyrimidines on the L strand). These two biases together also lead to a small bias for transitions over transversions (1.5-fold; a uniform transition bias alone will *not*, however, lead to a bias favoring purines or pyrimidines).

A priori, we would expect that when a substitution occurs, the type of change (e.g., $G \rightarrow A$, $C \rightarrow T$, etc.) would be independent of the strand on which it occurs. In particular, the relative substitution frequency (see footnote to Table 12) of $A \rightarrow G$ on the H strand should be approximately equal to that of $T \rightarrow C$ on the H strand since the latter is a $A \rightarrow G$ substitution on the complementary L strand. Similarly, $G \rightarrow A$ should equal $C \rightarrow T$ on the H strand. This does not appear to be true, however, for human mtDNA (Table 12). For example, the relative substitution frequency of $A \rightarrow G$ on the H strand (26.8%) is almost twice that of $A \rightarrow G$ on the L strand (14.6%). The probability that a given substitution will be a particular transition is thus not independent of the strand on which the substitution occurs. The finding of a predominance of $C \leftrightarrow T$ changes on the L strand, and $G \leftrightarrow A$ substitutions on the H strand, in comparisons among the cytochrome oxidase subunit II genes of rat, cow and human mtDNA (BROWN

TABLE 12
Relative substitution frequencies in human mtDNA

Substitution type		Relative substitution frequency ^a
H strand	L strand	
Transitions		
A → G	T → C	0.268
T → C	A → G	0.146
G → A	C → T	0.310
C → T	G → A	0.244
Transversions		
G → C	C → G	0.016
T → G	A → C	0.016

^a The expected proportion of the indicated base change in a random sequence (GOJOBORI, LI and GRAUR 1982), calculated as follows: $f_{ij} = P_{ij} (\sum_i \sum_{j \neq i} P_{ij})^{-1}$, where P_{ij} is the proportion of nucleotides of the i^{th} type changing to the j^{th} type ($i, j = A, G, C$ or T ; calculated from Table 3). For example, $P_{AG} = 13/214 = 0.061$ on the H strand.

and SIMPSON 1982) is consistent with this pattern and suggests that these biases may be general to coding as well as noncoding regions of the mtDNA molecule. It is of interest to note that a similar bias has been described in nuclear genes (VOGEL and KOPUN 1977; FITCH 1980; GOJOBORI, LI and GRAUR 1982).

The strand dependence of substitutions is qualitatively consistent with the observed density and base composition differences between complementary strands of animal mtDNA (reviewed by BROWN 1981). However, the equilibrium base composition predicted from the inferred substitution probabilities ($A = 0.31, G = 0.27, C = 0.17, T = 0.25$) is only roughly similar to that observed in the D-loop region of human mtDNA (mean for the seven sequences: $A = 0.24, G = 0.32, C = 0.14, T = 0.30$). This disparity is not too surprising considering the limited data from which the substitution probabilities were calculated and may indicate that our comparisons do not accurately reflect the substitution process and/or imply selection acting on base composition in this noncoding region.

In comparisons of an 896-base pair mtDNA coding sequence, BROWN *et al.* (1982) observed the transitional bias to vary from approximately 22- to fivefold between closely and more distantly related primates, respectively. Further comparisons of these primate sequences with mouse and cow mtDNA revealed a transition bias of only 1.6-fold. This apparent drop in the ratio of transitions to transversions over evolutionary time (5–80 million years) is probably a consequence of the transitional bias itself. If, in general, more than 96% of the base substitutions are transitions, transversions that do occur will usually be “fixed” in divergent lineages. That is, additional changes at these bases have a high probability of being transitions so that substitution back to the original nucleotide is unlikely. This process, over evolutionary time, will decrease the calculated transitional bias in comparisons of divergent sequences. An implication of this process is that quantitative methods for estimating divergence from pairwise sequence comparisons which rely on the observed proportion of transition and transversions (e.g., KIMURA’s 1981 K) will be biased due to a

serious underestimation of the actual proportion of transitions in the analysis of divergent mtDNA sequences.

In sharp contrast to the 32-fold bias toward transitional substitutions in mtDNA, transitions are generally found only one-half to two times as often as transversions in interspecific comparisons of nuclear genes (VOGEL and KOPUN 1977; VAN OOYEN *et al.* 1979; FITCH 1980; GOJOBORI, LI and GRAUR 1982). Although a high proportion of transitions (86%) has been observed in a comparison of two allelic human γ -globin sequences (SLIGHTOM, BLECHL and SMITHIES 1980), a large transition bias does not appear to be typical of all nuclear sequences. For example, two allelic sequences including a rat immunoglobulin gene that differ in sequence by only 1.8% show only a fourfold bias for transitions (SHEPPARD and GUTMAN 1981). Likewise, two allelic β -globin gene sequences from humans differ by an equal number of transitions and transversions (three of each over 1764 nucleotides; SPENCE *et al.* 1982).

The reason for the higher proportion of transitions in mtDNA evolution is unclear. The occurrence of the transitional bias in tRNA and protein-coding genes (BROWN *et al.* 1982; BROWN and SIMPSON 1982) as well as the noncoding sequences we have examined suggests that a significant portion of the bias may be due to the mutational process rather than selection against transversions. A higher frequency of transitions than transversions has been predicted for spontaneous mutations on the basis of chemical considerations and model building (TOPAL and FRESCO 1976). However, as was observed for nuclear genes by GOJOBORI, LI and GRAUR (1982), other predictions of TOPAL and FRESCO's model are in direct conflict with the observed data. For example, the relative substitution frequencies of transitions are not equal to each other (Table 12) as predicted nor is the observed strand bias predicted from their model. GOJOBORI, LI and GRAUR (1982) reveal that for nuclear DNA, the transition C:G \rightarrow T:A appears to be elevated in frequency due to the deaminative conversion of methylated C residues to T. If this mechanism was a major source of mutation in animal mtDNA, the relative substitution frequencies of G:C \rightarrow A:T plus C:G \rightarrow T:A should be significantly greater than T:A \rightarrow C:G plus A:T \rightarrow G:C. The difference, however, is small (55.4 vs. 41.4%), especially relative to the twofold difference seen for nuclear genes, and clearly could not account for the extreme transitional bias. In addition, there is no convincing evidence for base methylation in human mtDNA (DAWID 1974).

An attractive candidate for a source of mutational bias is the mtDNA polymerase itself. This enzyme has already been demonstrated to be approximately five times less accurate in DNA replication than the polymerase responsible for replicating nuclear DNA (KUNKEL and LOEB 1981), a difference that may contribute to the higher rate of substitution in mtDNA than observed for single copy nuclear DNA (BROWN and SIMPSON 1982). Whether the lower fidelity of replication characteristic of mtDNA polymerase is general or is specific to pathways that would lead to a transitional bias remains to be examined.

In agreement with FITCH (1980) and HOLMQUIST and PEARL (1980), our results also demonstrate that the rate and distribution of substitutions are not adequately described by current models which assume a single rate of substitution

for all sites. Within even the relatively small (899 nucleotide) region of the mtDNA genome we have examined in this study, the magnitude of variation in rates of substitution is significant. Some regions have been almost completely conserved for up to 80 million years (comparisons of human, mouse, rat and bovine mtDNA sequences; WALBERG and CLAYTON 1981; BIBB *et al.* 1981; ANDERSON *et al.* 1982). In sharp contrast, other regions (for example, near the H strand origin of replication) have diverged by as much as 15% in just a few hundred thousand years (comparisons among human sequences; time estimates from BROWN, GEORGE and WILSON 1979, and BROWN 1980).

In addition, the construction of a phylogenetic tree among the seven human sequences has allowed us to uncover a large number of multiple substitutions (at six of the 45 substitution sites). This high level of multiple substitution was unexpected between such closely related sequences under the assumption of uniform substitution rates. Five of the six multiple changes were parallel or convergent substitutions, consistent with the predominance of transitions. Evidence that these proposed convergent and parallel substitutions were not the product of recombination between divergent mtDNA molecules comes from the uniclonal and uniparental inheritance of mtDNA (detailed in the introduction) and the observation that the convergent or parallel changes appear among sequences that are also divergent at a large proportion of the 35 unique substitution sites. In addition, divergence at these unique sites is evenly distributed along these sequences, thus providing no evidence for recombination events. Whether this marked inequality of substitution rates and resulting high proportion of multiple substitutions is unique to the D-loop region we have examined or is typical of the whole mtDNA genome is unknown and must await the acquisition of additional human mtDNA sequences from other regions.

Thus, the rate of nucleotide substitution varies by several orders of magnitude in a noncoding region of the genome. For protein coding genes there appears to be some predictability concerning the rate of substitution depending upon the codon position; however, even these rates often vary greatly among gene families (e.g., KIMURA 1981; MIYATA, YASUNAGA and NISHIDA 1980; SHEPPARD and GUTMAN 1981). Our current ignorance of structure/function relationships and potential constraints across noncoding regions gives us even less basis for predicting rates of evolution in these regions. The modification of current population genetic models of variation and differentiation to incorporate this unpredictable diversity of substitution rates will provide a formidable challenge. Similarly, it will be important to take into account base composition and the significant transition bias in mtDNA, particularly for the estimation of sequence variability and rates of substitution among relatively divergent sequences.

Various lines of evidence, including restriction site variation and DNA-DNA hybridization, have suggested that mtDNA evolves at a rate five to ten times faster than single copy nuclear DNA (UPHOLT and DAWID 1977; BROWN, GEORGE and WILSON 1979). The apparently higher rate of transition observed for mtDNA would result in a higher proportion of convergent and parallel substitutions in mtDNA than in nuclear DNA. These multiple substitutions would not have been detected in the pairwise comparisons from which the relative rates were

initially calculated. Thus, the disparity between average rates of single copy nuclear and mitochondrial DNA evolution may be significantly greater than first suspected.

The human mtDNA sequences we have examined in this study represent one of the first "population" samples of nucleotide sequences available. Our results demonstrate that comparisons among divergent sequences can give a significantly biased view of the substitution process. In addition, the construction of a phylogeny for the sequences has afforded a unique view of the frequency of parallel and convergent substitution and supports the hypothesis that, in contrast to nuclear genes, recombination plays very little if any role in the generation of mtDNA sequence diversity in animals. Although it will be more difficult to discriminate convergent and parallel substitutions from recombination events in nuclear DNA sequences, the insight provided by the examination of additional intraspecific samples of sequences should, like those in our current study, significantly contribute to our understanding of the range of sequence variations in natural populations and the nature of sequence evolution.

We thank S. ANDERSON, J. C. AVISE, C. W. BIRKY, JR., W. M. BROWN, D. CLAYTON, M. T. CLEGG, T. GOJOBORI, N. KAPLAN, R. A. LANSMAN and A. C. WILSON for communicating their results prior to publication and our colleagues, particularly N. KAPLAN, C. H. LANGLEY and W. M. BROWN, for helpful discussions and constructive criticism. The comments and suggestions of B. S. WEIR, M. T. CLEGG, G. R. CARMODY and two reviewers were also helpful in strengthening our presentation. L. GARDNER's care and patience in the preparation of this manuscript are appreciated.

LITERATURE CITED

- ADAMS, J. and E. D. ROTHMAN, 1982 Estimation of phylogenetic relationships from DNA restriction patterns and selection of endonuclease cleavage sites. *Proc. Natl. Acad. Sci. USA* **79**: 3560-3564.
- ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. DE BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICH, 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.
- AVISE, J. C., C. GILBIN-DAVIDSON, J. LAERM, J. C. PATTON and R. A. LANSMAN, 1979 Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. *Proc. Natl. Acad. Sci. USA* **76**: 6694-6698.
- BIBB, M. J., R. A. VAN ETEN, C. T. WRIGHT, M. W. WALBERG and D. A. CLAYTON, 1981 Sequence and gene organization of mouse mitochondrial DNA. *Cell* **26**: 167-180.
- BIRKY, C. W., JR., T. MARUYAMA and P. A. FUERST, 1982 Population and evolutionary genetic theory for genes in mitochondria and chloroplasts. *Genetics*, in press.
- BREIMAN, L., 1973 *Statistics with a View towards Application*. Houghton Mifflin, Boston.
- BROWN, A. H. D. and M. T. CLEGG, 1983 Analysis of variation in related sequences. In: *Statistical Analysis of DNA Sequence Data*, Edited by B. S. WEIR. Marcel Dekker, New York, in press.
- BROWN, G. G., F. J. CASTORA, S. C. FRANTZ and M. V. SIMPSON, 1981 Mitochondrial DNA polymorphism: evolutionary studies on the genus *Rattus*. *Ann. NY Acad. Sci.* **361**: 135-153.
- BROWN, G. G. and M. V. SIMPSON, 1981 Intra- and interspecific variation of the mitochondrial

- genome in *Rattus norvegicus* and *Rattus rattus*: restriction enzyme analysis of variant mitochondrial DNA molecules and their evolutionary relationships. *Genetics* **97**: 125-143.
- 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. USA* **79**: 3246-3250.
- BROWN, W. M., 1980 Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc. Natl. Acad. Sci. USA* **77**: 3605-3609.
- BROWN, W. M., 1981 Mechanisms of evolution in animal mitochondrial DNA. *Ann. NY Acad. Sci.* **361**: 119-134.
- 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. and H. M. GOODMAN, 1979 Quantification of intrapopulation variation by restriction endonuclease analysis of human mitochondrial DNA. pp. 485-499. In: *Extrachromosomal DNA*, Edited by D. J. CUMMINGS, P. BORST, I. B. DAWID, S. M. WEISSMAN and C. F. FOX. Academic Press, New York.
- 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.
- CHAPMAN, R. W., J. C. STEPHENS, R. A. LANSMAN and J. C. AVISE, 1982 Models of mitochondrial DNA transmission genetics and evolution in higher eucaryotes. *Genet. Res.* **40**: 41-57.
- DAWID, I. B., 1974 5-Methylcytidylic acid: absence from mitochondrial DNA of frogs and HeLa cells. *Science* **184**: 80-81.
- DE FRANCESCO, L., G. ATTARDI and C. M. CROCE, 1980 Uniparental propagation of mitochondrial DNA in mouse-human cell hybrids. *Proc. Natl. Acad. Sci. USA* **77**: 4079-4083.
- DENARO, M., H. BLANC, M. J. JOHNSON, K. H. CHEN, E. WILMSEN, L. L. CAVALLI-SFORZA and D. C. WALLACE, 1981 Ethnic variation in *HpaI* endonuclease cleavage patterns of human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **78**: 5768-5772.
- DROUIN, J., 1980 Cloning of human mitochondrial DNA in *Escherichia coli*. *J. Mol. Biol.* **140**: 15-34.
- ENGELS, W. R., 1981 Estimating genetic divergence and genetic variability with restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **78**: 6329-6333.
- EWENS, W. J., R. S. SPIELMAN and H. HARRIS, 1981 Estimation of genetic variation at the DNA level from restriction endonuclease data. *Proc. Natl. Acad. Sci. USA* **78**: 3748-3750.
- FERRIS, S. D., W. M. BROWN, W. S. DAVIDSON and A. C. WILSON, 1981 Extensive polymorphism in the mitochondrial DNA of apes. *Proc. Natl. Acad. Sci. USA* **78**: 6319-6323.
- FERRIS, S. D., A. C. WILSON and W. M. BROWN, 1981 Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **78**: 2432-2436.
- FITCH, W. M., 1977 On the problem of discovering the most parsimonious tree. *Am. Nat.* **111**: 223-257.
- FITCH, W. M., 1980 estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNAs. *J. Mol. Evol.* **16**: 153-209.
- 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.
- GILES, R. E., H. BLANC, H. M. CANN and D. C. WALLACE, 1980 Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **77**: 6715-6719.
- 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.

- GOJOBORI, T., W.-H. LI and D. GRAUR, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**: 360-369.
- GREENBERG, B. D., J. E. NEWBOLD and A. SUGINO, 1982 Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. *Gene*, in press.
- HAUSWIRTH, W. W. and P. J. LAIPIS, 1982 Mitochondrial DNA polymorphism in a maternal lineage of Holstein cows. *Proc. Natl. Acad. Sci. USA* **79**: 4686-4690.
- HAYASHI, J.-I., H. YONEKAWA, O. GOTOH, Y. TAGASHIRA, K. MORIWAKI and T. H. YOSIDA, 1979 Evolutionary aspects of variant types of rat mtDNAs. *Biochem. Biophys. Acta* **564**: 202-211.
- HOLMQUIST, R. and D. PEARL, 1980 Theoretical foundations for quantitative paleogenetics. Part III: The molecular divergence of nucleic acids and proteins for the case of genetic events of unequal probability. *J. Mol. Evol.* **16**: 211-267.
- HUDSON, R. R., 1982 Estimating genetic variability with restriction endonucleases. *Genetics* **100**: 711-719.
- HUTCHISON, C. A., III, J. E. NEWBOLD, S. S. POTTER and M. H. EDGEELL, 1974 Maternal inheritance of mammalian mitochondrial DNA. *Nature* **251**: 536-538.
- JAKOVIC, S., J. CASEY and M. RABINOWITZ, 1975 Sequence homology between mitochondrial DNAs of different eukaryotes. *Biochemistry* **14**: 2043-2050.
- JEFFREYS, A. J., 1981 Recent studies of gene evolution using recombinant DNA. pp. 1-48. In: *Genetic Engineering* Vol. 2, Edited by R. WILLIAMSON. Academic Press, New York.
- KAPLAN, N., 1983 Statistical analysis of restriction enzyme map data and nucleotide sequence data. In: *Statistical Analysis of DNA Sequence Data*, Edited by B. S. WEIR. Marcel Dekker, New York.
- KAPLAN, N. and K. RISK, 1981 An improved method for estimating sequence divergence of DNA using restriction endonuclease mapping. *J. Mol. Evol.* **17**: 156-162.
- Kaplan, N. and K. Risk, 1982 A method for estimating rates of nucleotide substitution using DNA sequence data. *Theor. Pop. Biol.* **21**: 318-328.
- KIMURA, M. 1981 Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78**: 454-458.
- KUNKEL, T. A. and L. A. LOEB, 1981 Fidelity of mammalian DNA polymerases. *Science* **213**: 765-767.
- LANSMAN, R. A., J. C. AVISE, C. F. AQUADRO, J. F. SHAPIRA and S. W. DANIEL, 1982 Extensive genetic variation in mitochondrial DNA's among geographic populations of the deer mouse, *Peromyscus maniculatus*. *Evolution*, in press.
- LANSMAN, R. A., J. C. AVISE and M. D. HUETTEL, 1983 Critical experimental test of the possibility of "paternal leakage" of mitochondrial DNA. *Proc. Natl. Acad. Sci. USA*, in press.
- MIYATA, T., T. YASUNAGA and T. NISHIDA, 1980 Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc. Natl. Acad. Sci. USA* **77**: 7328-7332.
- NEI, M. and W.-H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**: 5269-5273.
- NEI, M. and F. TAJIMA, 1981 DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**: 145-163.
- OJALA, D., S. CREWS, J. MONTOYA, R. GELFAND and G. ATTARDI, 1981 A small polyadenylated RNA (7 S RNA), containing a putative ribosome attachment site, maps near the origin of human mitochondrial DNA replication. *J. Mol. Biol.* **150**: 303-314.
- OKADA, Y., G. STREISINGER, J. OWEN, J. NEWTON, A. TSUGITA and M. INOUE, 1972 Molecular basis of a mutational hot spot in the lysozyme gene of bacteriophage T4. *Nature* **236**: 338-341.

- POTTER, S. S., J. E. NEWBOLD, C. A. HUTCHISON, III and M. H. EDGELL, 1975 Specific cleavage analysis of mammalian mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **72**: 4496-4500.
- SHAH, D. M. and C. H. LANGLEY, 1979 Inter- and intraspecific variation in restriction maps of *Drosophila* mitochondrial DNA. *Nature* **281**: 696-699.
- SHEPPARD, H. W. and G. A. GUTMAN, 1981 Allelic forms of rat κ chain genes: evidence for strong selection at the level of nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78**: 7064-7068.
- SLIGHTOM, J. L., A. E. BLECHL and O. SMITHIES, 1980 Human fetal $\zeta\gamma$ - and $\alpha\gamma$ -globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21**: 627-638.
- SPENCE, S. E., R. G. PERGOLIZZI, M. DONOVAN-PELUSO, K. A. KOSCHE, C. S. DOBKIN and A. BANK, 1982 Five nucleotide changes in the large intervening sequence of a β globin gene in a β^+ thalassemia patient. *Nucl. Acids Res.* **10**: 1283-1294.
- STREISINGER, G., Y. OKADA, J. EMRICH, J. NEWTON, A. TSUGITA, E. TERZAGHI and M. INOUE, 1966 Frameshift mutations and the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31**: 77-84.
- TAJIMA, F. and M. NEI, 1982 Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *J. Mol. Evol.* **18**: 115-120.
- TAKAHATA, N. and T. MARUYAMA, 1981 A mathematical model of extranuclear genes and the genetic variability maintained in a finite population. *Genet. Res.* **37**: 291-302.
- TOPAL, M. D. and J. R. FRESCO, 1976 Complementary base pairing and the origin of substitution mutations. *Nature* **263**: 285-289.
- 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.
- VAN OUYEN, A., J. VAN DEN BERG, N. MANTEL and C. WEISSMANN, 1979 Comparison of total sequence of a cloned rabbit β -globin gene and its flanking regions with a homologous mouse sequence. *Science* **206**: 337-344.
- VOGEL, F. and M. KOPUN, 1977 Higher frequencies of transitions among point mutations. *J. Mol. Evol.* **9**: 159-180.
- 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. *Nucl. Acids Res.* **9**: 5411-5421.

Corresponding editor: B. WEIR