Copyright © 1986 by the Genetics Society of America

VARIABILITY OF EVOLUTIONARY RATES OF DNA

JOHN H. GILLESPIE

Department of Genetics, University of California, Davis, California 95616

Manuscript received October 10, 1985
Revised copy accepted May 5, 1986

ABSTRACT

A statistical analysis of DNA sequences from four nuclear loci and five mitochondrial loci from different orders of mammals is described. A major aim of the study is to describe the variation in the rate of molecular evolution of proteins and DNA. A measure of rate variability is the statistic R, the ratio of the variance in the number of substitutions to the mean number. For proteins, R is found to be in the range $0.16 < R < 35.55$, thus extending in both directions the values seen in previous studies. An analysis of codons shows that there is a highly significant excess of double substitutions in the first and second positions, but not in the second and third or first and third positions. The analysis of the dynamics of nucleotide evolution showed that the ergodic Markov chain models that are the basis of most published formulas for correcting for multiple substitutions are incompatible with the data. A bootstrap procedure was used to show that the evolution of the individual nucleotides, even the third positions, show the same variation in rates as seen in the proteins. It is argued that protein and silent DNA evolution are uncoupled, with the evolution at both levels showing patterns that are better explained by the action of natural selection than by neutrality. This conclusion is based primarily on a comparison of the nuclear and mitochondrial results.

THE concept of a molecular clock grew out of the observation that proteins appear to evolve at a nearly constant rate (ZUCKERKANDL and PAULING 1965; WILSON, CARLSON and WHITE 1977). However, as early as 1971, OHTA and KIMURA pointed out that the rates of evolution are not constant, but rather, they vary significantly from lineage to lineage. This observation has subsequently been verified by a number of people (*e.g.*, LANGLEY and FITCH 1974; KIMURA 1983). A common statistic that quantifies the variability in rates of evolution is R, defined as the ratio of the variance in the number of substitutions in a lineage to the mean number. For proteins, R is usually in the range $1.0 < R < 3.4$ (LANGLEY and FITCH 1974; KIMURA 1983; GILLESPIE 1984b, among others).

There are two very different interpretations of these estimates of R. KIMURA (1983) claims that, as R is close to one, it suggests that evolutionary rates are nearly constant and that the events of molecular evolution may be approximated by a Poisson process (for which $R = 1$). This interpretation is also used to support the neutral allele theory, since the substitution process for this theory is close to a Poisson process (GILLESPIE and LANGLEY 1979; WATTERSON 1982a,b). A second interpretation is that inferences about the variance in evolutionary rates from the observed values of R are severely biased toward one (GILLESPIE 1984b, 1986b). When this bias is accounted for, the fact that we can measure R values as large as two or three suggests that the real variance in evolutionary rates might be very large—so large, in fact, that molecular evolution may well be episodic, with bursts of substitutions separated by long periods with no substitutions.

In this paper an effort will be made to examine DNA sequences for variability in rates in a manner analogous to the studies on proteins. Of particular interest will be a comparison of the R values for silent *vs.* replacement substitutions. This comparison should provide some insight into the forces responsible for molecular evolution. As will become apparent, the analysis of DNA data is considerably more complex because of the difficulties in correcting for multiple substitutions at a site. Most of the published correction formulas are inadequate to account for the patterns in the data for sequences that differ in 30% or more of the bases. Nonetheless, by using various parametric and nonparametric statistical techniques, some surprising patterns emerge from the data. The analyses will be presented in a hierarchical fashion, beginning at the top with the entire protein, then with codons and, finally, with the individual nucleotides.

THE DNA SEQUENCES

The study of the variance in the rate of evolution is subject to fewer biases if the species that are used come from a star phylogeny; that is, if the species arose from a radiation in the remote past that occupied a short period of time relative to the time back to the radiation. In this setting, the mysteries of tree construction are avoided with their largely unknown sources of bias for estimating variance of rates. The radiation that gave rise to the modern orders of mammals is the most useful example of a radiation that closely approximates a star phylogeny. In this study, all of the sequences that will be used for estimating variances in rates use different orders of mammals, each being a member of one leg of the star phylogeny. KIMURA (1983) was the first to recognize the value of star phylogenies for the estimation of variance in rates. His statistical techniques form the basis of the approach that will be used in this paper.

The DNA sequences that were used in this study are listed in Table 1. The sequences were downloaded from the Intelligenetics database to a Symmetric 32016-based computer for subsequent local analysis. Loci were chosen if there were sequences available from at least three orders of mammals and if they came from coding regions that are all of the same length in the different species. The limitation to sequences of the same length avoids any biases that may be introduced from procedures that align sequences. The loci that fit these two criteria are called the "standard" loci.

In addition to the standard loci, it proved necessary to look at pairs of sequences that were either more similar or less similar than the standard loci.

TABLE 1

DNA sequences

Species	Locus	Name	Splice 1	Splice 2	Splice 3
Nuclear loci					
Human	*hba*	HUMHBA5	135:229	347:551	692:817
Mouse	*hba*	MUSHBA	405:499	622:826	961:1086
Goat	*hba*	GOTHBAI	917:1011	1120:1324	1429:1554
Rabbit	*hba*	RABHBA	37:462		
Chicken^{o}	*hba*	CHKHBADA2	337:431	563:767	877:1002
Duck^{o}	*hba*	DUKHBADA2	367:461	612:816	921:1046
Human	*hbb*	HUMHBB	276:359	490:711	1562:1687
Mouse	*hbb*	MUSHBBMAJ	242:324	441:663	1317:1442
Bovine	*hbb*	BOVHBB	281:363	492:714	1613:1738
Rabbit	*hbb*	RABHBB1A1	493:575	702:924	1498:1623
Chicken^{o}	*hbb*	CHKHBBM	52:492		
Human	*ins*	HUMINS1	2424:2610	3397:3539	
Dog	*ins*	DOGINS	324:510	775:917	
Rat	*ins*	RATINS1	287:616		
Monkey^{o}	*ins*	MNKINS	60:389		
Guinea pig^{o}	*ins*	GPIINS	339:525	1139:1281	
Human	*prl*	HUMPRL	119:685		
Rat	*prl*	RATPRLSDM	160:726		
Bovine	*prl*	BOVPRL	188:754		
Mitochondrial loci					
Human	*cytox1*	HUMMT	5904:7442		
Mouse	*cytox1*	MUSMT	5328:6866		
Bovine	*cytox1*	BOVMT	5687:7225		
Rat^{o}	*cytox1*	RATMTCYOS	434:1972		
Human	*cytox2*	HUMMT	7586:8266		
Mouse	*cytox2*	MUSMT	7013:7693		
Bovine	*cytox2*	BOVMT	7374:8054		
Rat^{o}	*cytox2*	RATMTCYOS	2117:2797		
Human	*cytox3*	HUMMT	9207:9989		
Mouse	*cytox3*	MUSMT	8607:9389		
Bovine	*cytox3*	BOVMT	8970:9752		
Rat^{o}	*cytox3*	RATMTCYOS	3710:4492		
Human	*atp6*	HUMMT	8527:9204		
Mouse	*atp6*	MUSMT	7927:8604		
Bovine	*atp6*	BOVMT	8290:8967		
Rat^{o}	*atp6*	RATMTCYOS	3030:3707		
Human	*cytb*	HUMMT	14747:15883		
Mouse	*cytb*	MUSMT	14139:15275		
Bovine	*cytb*	BOVMT	14514:15650		
Rat^{o}	*cytb*	RATMTCYBT	312:1448		

The abbreviations for the loci are as follows: α-hemoglobin, *hba*; β-hemoglobin, *hbb*; insulin, *ins*; prolactin, *prl*; cytochrome oxidase n, *cytoxn* (n = 1–3); atpase 6, *atp6*; cytochrome b, *cytb*. The locus names are the ones used by Genbank. The splices are the numbers of the nucleotides that were used. The loci marked by a superscript "o" are from the 'other' set of loci; all of the others are from the "standard" set.

TABLE 2

Protein analysis

Locus	$\hat{r}$	Mean no. of episodes		Mean no. of substitutions per episode	
		Gamma	Two-state	Gamma	Two-state
Nuclear loci					
prl	11.90***	8.74	7.06	4.40	5.45
hba	0.95	NA	NA	NA	NA
ins	3.07*	5.30	9.46	1.85	1.03
hbb	5.31***	5.58	6.68	2.58	2.15
Mitochondrial loci					
atp6	0.60	NA	NA	NA	NA
cytb	6.42***	14.10	15.16	2.92	2.71
cytox1	3.41*	11.13	18.13	1.97	1.21
cytox2	35.55***	2.87	1.61	9.68	17.28
cytox3	0.16	NA	NA	NA	NA

Summary of the statistical analysis of the translated DNA sequences. The entries under $\hat{r}$ are significant at the 5% level (*) and at the 0.5% level (***). NA indicates that these calculations are not appropriate for proteins for which the estimated value of R, $\hat{r}$, is less than one.

The loci used in this part of the study are referred to as "other" loci and are also listed in Table 1.

PROTEINS

The major goal of the protein analysis was to translate the sequences and to estimate the value of R, the ratio of the variance in the number of substitutions per lineage to the mean number. Since the sequences all come from star phylogenies, the technique for estimating R is a simple variant of the method described by KIMURA (1983). In this method we suppose that there have been $N_i(t)$ substitutions on a lineage of length t leading to the ith species. In this context, R is defined as

$$R(t) = \mathrm{Var}\{N_i(t)\}/E\{N_i(t)\} = \sigma^2/\mu,$$

where we have taken the opportunity to define μ and σ^2. The method used to estimate R is taken from GILLESPIE (1986b). The estimated values of $R,\hat{r}$, are given in Table 2. There are several striking aspects of the results. The range of the estimates is $0.16 < \hat{r} < 35.5$. In previous analyses using only star phylogenies (KIMURA 1983; GILLESPIE 1984b, 1986a,b), the observed range of values is $1.1 < \hat{r} < 3.5$. Thus, the range has been greatly extended in both directions by this new data.

KIMURA (1983) argued that $(n - 1)\hat{r}$ should be χ^2 distributed with $(n - 1)$ degrees of freedom if the substitution process, $N_i(t)$, is a Poisson process. Using this statistic, we see in Table 2 that none of the three values of $\hat{r}$ that are less than one allow rejection of the Poisson process, whereas all six of the values that are greater than one allow rejection of the Poisson process.

Of the nine proteins considered in Table 2, four exhibit $\hat{r}$ values that are larger than any seen previously: prolactin ($\hat{r}$ = 11.9), β-hemoglobin ($\hat{r}$ = 5.31), cytochrome b ($\hat{r}$ = 6.42) and cytochrome oxidase 2 ($\hat{r}$ = 35.5). The value of 35 for cytochrome oxidase 2 stems from the fact that the human enzyme differs from that of the mouse and cow by 65 and 61 amino acids, respectively, whereas mouse and bovine differ by only 21. If we take these raw numbers and crudely consider them to be the actual number of substitutions that occurred between the three pairs of species, then we are led through a little algebra to the conclusion that 53 substitutions occurred on the lineage leading to humans from the mammal radiation, whereas only 12 occurred on the lineage leading to the mouse and nine occurred on the lineage leading to the cow. Thus, within these three species, there is a sixfold difference in rates.

For prolactin, the high value of $\hat{r}$ is attributable to a crude (noncorrected) estimate of 22 substitutions on the lineage leading to humans, 25 substitutions on the lineage to the cow, and 48 substitutions on the lineage leading to the rat. These rates exhibit a twofold variation, considerably less than in the previous case. The other examples of large values of $\hat{r}$ will yield somewhat less spectacular variation in rates.

In a series of papers, GILLESPIE (1984b, 1986a,b) argued that the variation in the numbers of substitutions per lineage must ultimately be attributable to variation in the rates of substitutions. However, efforts to estimate the variation in the rates of substitutions led to the conclusion that the variance is so large that the substitution process is best viewed as an episodic process in which bursts of substitutions are followed by long periods with no substitutions. A method was described for estimating the mean number of episodes in a lineage and the mean number of substitutions per episode for two versions of the molecular clock. One of these, the two-state clock, is a clock that fluctuates between two rates: zero and a very high fixed rate. The other one, the gamma clock, is a clock that alternates between a rate of zero and a gamma-distributed rate. The estimates for the means for these two clocks are given in Table 2. As noticed in other data, the agreement between the clocks is quite good except for the case of cytochrome oxidase 2, for which the two clocks give very different answers.

The new data are similar to that analyzed in the previous papers in that, for all cases except cytochrome oxidase 2 and prolactin, there tends to be a mean number of substitutions per episode in the range one to three. Cytochrome oxidase once again stands out as being conspicuously different from the other proteins that have been examined.

CODONS

The estimates of R for silent and replacement substitutions, using the correction formulas of MIYATA *et al.* (1982), are presented in Table 3. This correction formula fails for some of the mitochondrial silent sites. For these sites the formula given by MIYATA and YASUNGA (1980) is used. The estimated values of R for replacement sites are similar to those for entire proteins. For silent sites $0.71 < \hat{r} < 19.42$. Six of nine of the values of $\hat{r}$ are greater than three, just as with the replacement values.

TABLE 3

Codon analysis

	$\hat{r}$		No. of significant χ^2					
Locus	aa subs.	s subs.	Total	2 and 3	1 and 3	1 and 2	1, 2 and 3	Q
Nuclear loci								
prl	14.28	9.67	1/3	0/3	0/3	1/3	0/3	1.79
hba	0.94	19.42	5/6	3/6	0/6	4/6	2/6	2.65
ins	3.24	4.01	2/3	1/3	0/3	0/3	1/3	2.19
hbb	7.52	4.06	5/6	0/6	0/6	5/6	1/6	3.82
Mitochondrial loci								
atp6	0.24	7.74	3/3	1/3	1/3	3/3	1/3	2.40
cytb	11.39	3.22	3/3	0/3	0/3	3/3	0/3	2.69
cytox1	4.98	0.71	2/3	0/3	0/3	2/3	1/3	4.48
cytox2	34.14	1.22	1/3	0/3	0/3	1/3	0/3	2.73
cytox3	0.49	1.83	3/3	0/3	0/3	2/3	0/3	3.04

The values of $\hat{r}$ are calculated using the correction formula of MIYATA *et al.* (1982), except for silent substitutions for the mitochondria, for which the formula of MIYATA and YASUNAGA (1980) is used. The results are reported separately for amino acid substitution (aa) and silent substitution(s). The results for the χ^2 tests are given in the form (number of significant tests at the 1% level)/(number of tests). Q is the observed number of differences in the first two positions of a codon divided by the expected number.

The analysis of codons can be carried a little deeper by looking for patterns within the codons. A χ^2 test was used to test for interactions in the substitutions at different positions within a codon. First, for each pair of sequences the frequency of base differences in the first, second and third sites was calculated. Next, the probability of a codon differing in from zero to three sites in all of the eight possible configurations was predicted using the results of the first calculation and the assumption that differences at the sites are independent. This provides the expected frequencies for the χ^2 test. The results of the test are presented in Table 3. Of the 33 pairwise tests, 25 yielded a significant χ^2 value (at the 1% level). In interpreting this result it should be kept in mind that the 33 χ^2 tests are not independent of one another, because they reuse the same sequences. Nonetheless, the pattern is obvious: there is a high degree of interaction in the probability of substitutions at different sights.

To analyze this further, the χ^2 value was partitioned into interactions between positions 1 and 2, 2 and 3, and 1 and 3, as well as an interaction of all three positions. The results of the partitioning are also presented in Table 3. It is clear from this table that most of the significant interactions are between sites 1 and 2. When the predicted codon differences are compared to the expected, it is seen that, in all cases, the significant interaction between sites 1 and 2 is due to an excess of pairs of codons with differences in both sites one and two over what is expected under the assumption of no interaction. This pattern was observed by FITCH and MARGOLIASH (1966) using protein sequence data. The fact that we do not see the same level of significance for sites 2 and 3 suggests that the effect may not be due to localized mutational events.

FITCH (1971) attributed the excess of double substitutions to site-specific variation in the rate of evolution. His particular model assumes that there are a group of sites, called covarions, that are experiencing all of the substitutions, and another group of sites that are experiencing no substitutions. HOLMQUIST *et al.* (1983) have argued that the distribution of rates per site is closer to a negative binomial or exponential distribution than to the uniform distribution of the covarion model. These alternatives may be examined in light of our data. Our goal is to obtain an analytic expression for the ratio of the observed to expected number of double substitutions, under the assumption that the rate of substitution varies with the site, and to use this result to describe the nature of the site-specific variation in rates.

Let a_i be the probability that the first site in the codon is different when two species are compared. The probability that a randomly drawn first site is different is $\bar{a} = n^{-1}\Sigma a_i$, where the summation is taken over all of the sites. Similarly, the probability that a second site is different may be written as $\bar{b} = n^{-1}\Sigma b_i$. The "predicted" probability that the first two sites are different, under the assumption of independence of sites, is $\bar{a}\cdot\bar{b}$. The "observed" probability that the first two sites are different is the average of the a_ib_i or $\overline{ab} = n^{-1}\Sigma a_ib_i$. In general, $\bar{a}\cdot\bar{b}$ does not equal $\overline{ab}$. Now reparameterize the model, such that $a_i = r_ia$ and $b_i = r_ib$ where $\Sigma r_i = 1$. This restriction on the model from $2n$ to $n + 1$ parameters assumes that there is a codon-specific factor, r_i, and two position factors, a and b, that determine the probability that a difference is observed at a position within the *i*th codon. Under this parameterization the "observed" number of double differences over the "expected" number, $Q = \overline{ab}/\bar{a}\cdot\bar{b}$, is $n\Sigma r_i^2$, where n is the number of sites.

As a simple example, suppose the covarion model applies and that there are c covarions. If so, then $r_i = 1/c$ if the *i*th site is a covarion, else $r_i = 0$. For this model, $Q = n/c$. For β-hemoglobin Q = 3.82, n = 144 codons, giving c = 37.7 covarions. However, there is little justification for using such a simple model for the r_i. HOLMQUIST *et al.* (1983) have argued from actual sequence data that the codon-specific rates of β-hemoglobin are geometrically distributed. There are two problems with using this observation to argue that the r_i are also geometrically distributed. First, if each site has a characteristic rate, then the r_i are more properly viewed as fixed quantities, rather than as the distribution of some random quantity. In the HOLMQUIST paper there appears to be some confusion over this point. Second, our definition of the r_i refers to the probability of a site being different, rather than being a rate of substitution. The connection between these two is, in general, quite complex. If we blindly ignore these problems, then for β-hemoglobin the geometric distribution applies to 139 variable sites with a parameter of $P = 0.89$ [see HOLMQUIST *et al.* (1983) for these values] giving $Q = 8.49$. This is somewhat higher than the value of 3.82 seen in our data. Assuming that the difference is real, it is not unexpected. Recall that the genetic code is structured such that double substitutions, on the average, result in amino acid changes that chemically are rather drastic. MIYATA (1982), among others, have shown that single base substitutions that cause large chemical differences are underrepresented in the data. If this were to apply to double substitutions as well, then the fact that the observed Q is less than predicted by the HOLMQUIST data is to be expected. This argument is full of pitfalls, but does point out a very intriguing direction for further work.

NUCLEOTIDES

When we move down to the level of the nucleotides the analyses become much more suspect because of the problems associated with correcting for multiple substitutions at a site. For proteins, the relatively low rates of substitutions and the fact that amino acids can exist in 20 different states gives some confidence that multiple substitutions can be adequately accounted for. For nucleotides, particularly at the third position, there are many more substitutions, and each nucleotide site can be in one of only four states. There are many formulas in the literature that are claimed to correct for multiple substitutions [see KAPLAN (1983) and TAVARE (1986) for a review of some of them]. All of the published formulas are based on ergodic Markov chain models. The literature on correcting for multiple substitutions is vast, but very few attempts have been made to test whether the assumptions of the correction formulas are compatible with the data. A notable exception is the recent work by TAVARE (1986). He examined a small number of sequences and noted that they did not appear to meet the ergodic assumption as tested by two different χ^2 tests: one for base frequencies, the other for symmetry. We shall begin this section by applying these tests to our data set.

The first test involves comparing pairs of sequences to see if they have the same frequencies of the four bases. This is accomplished by using a simple 4 × 2 contingency analysis. Significant results were only found for the third position where 14 of 33 tests yielded significant χ^2 values at the 1% level. No significance was found for the first two positions.

A natural question to ask is whether the significance of the results for the third position is due to a real difference in the dynamics of the third sites or whether it is due to the fact that the third position has experienced more substitutions and, thus, the test has more power for detecting the difference. To explore this thought we shall use a measure for the difference in base frequency between two sequences and shall examine the behavior of this measure as a function of the fraction of sites that differ between a pair of sequences. If the frequencies of the four bases in one sequence are p_i, i = A, G, C, T and for the other sequence are q_i, then an informative measure would be

$$\delta_{bf} = \sum_i (p_i - q_i)^2 / \sum_i [2\hat{p}_i(1 - \hat{p}_i)/n],$$

where n is the number of nucleotides used in the comparison, $\hat{p}_i$ is the true frequency of the ith base for both sequences, and the summation is over the four bases. The denominator in this expression will be recognized as the expectation of the numerator if the two sequences are independent. Therefore,

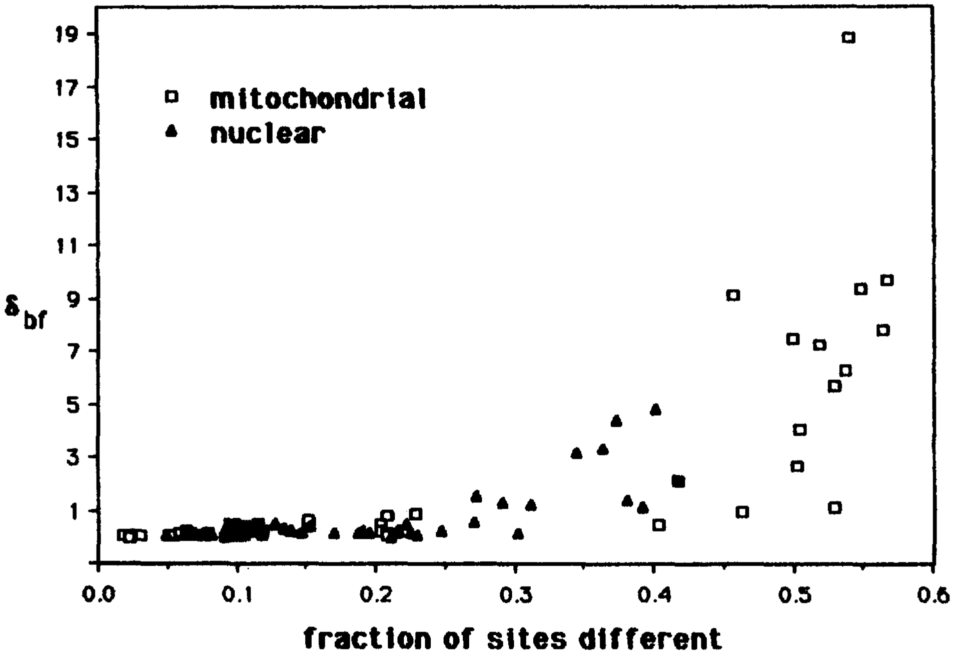

FIGURE 1.—The relationship between the fraction of sites different and the statistic δ_{bf} for the "standard" set of loci.

as the two sequences diverge, the expected value of δ_{bf} will change from zero to one. Unfortunately, we cannot use this measure directly because we cannot know the $\hat{p}_i$ with absolute certainty. However, the behavior will be almost what we desire if we set $\hat{p}_i = (p_i + q_i)/2$.

Figure 1 presents a plot of δ_{bf} as a function of the fraction of bases that are different for all pairs of sequences in the standard set of loci. As can be seen, when the fraction of bases that are different exceeds about 25%, δ_{bf} begins a steep rise well past one, the value where it should level off if the process were ergodic. In this figure the points for the first, second and third positions appear as separate clumps, and it is easily seen that the rise past one is due to the third positions of the nuclear and mitochondrial loci. Figure 2 presents the same plot for various pairs from the set of "other" loci matched to selected members of the standard set. In this figure, the first, second and third positions are intermixed because of the different times separating the pairs of loci. The three positions appear to exhibit the same relationship between the fraction of sites that are different and δ_{bf} as do the loci from the standard set. Thus, despite the paucity of suitable data, it nonetheless appears that all three positions are in violation of the ergodic assumption that is the basis of essentially all published correction formulas. Furthermore, it appears that, although the rate of change of the three positions is very different, the aspects of their evolution described by δ_{bf} appear similar.

Another test for peculiar behavior of nucleotides is a symmetry test suggested by TAVARE (1986). The basis of this test is the simple idea that if we

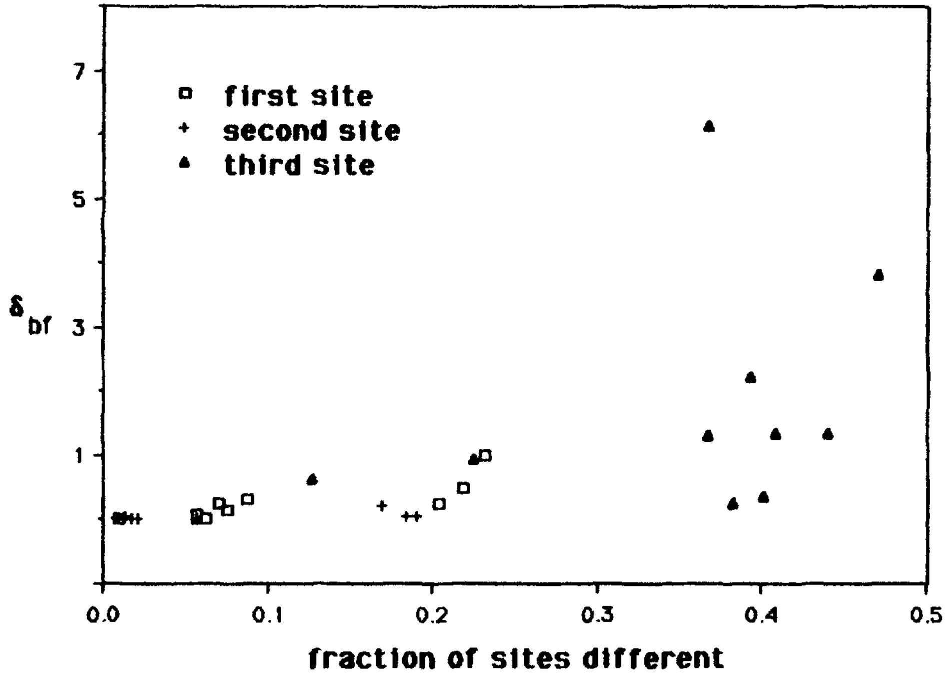

FIGURE 2.—The relationship between the fraction of sites different and the statistic δ_{bf} for the "other" set of loci.

observe, say, base A in species one and T in species two at a certain number of sites, then we should see about the same number of occurences of T in species one and A in species two. When Tavare's symmetry χ^2 was applied to the standard, only the third positions exhibited significant χ^2 values, and did so for 15 of 33 tests.

These two tests indicate that nucleotide dynamics do not conform to the assumptions of the published correction formulas. Therefore, it is difficult to see how we can estimate the value of R with any confidence. As an initial attempt we shall pick one of the many available correction formulas and proceed as if it were providing a reasonable insight into the variability of rates. The formula that will be used is due to LANAVE *et al.* (1984). It only assumes that the base frequency dynamics conform to a reversible Markov process. It is more general than most of the published formulas, although it does share the assumptions that the process is Markovian and ergodic. I have compared the inferred number of substitutions predicted by this formula to others and find that they all give very similar results unless the frequency of bases that differ between two sequences exceeds about 25%, in which case the formulas give very different answers. The LANAVE formula actually fails in many cases to even provide an answer because of the internal inconsistencies of the data once sequences differ by more than 25%. For this reason we should be highly suspicious of any results for sequences that differ in more than 25% of the positions.

TABLE 4

Nucleotide analysis

Locus	$\hat{r}$			Bootstrap		
	1	2	3	1	2	3
Nuclear loci						
prl	1.42 (0.89)	15.87 (9.92)	9.80 (3.80)	0.35	0.000	0.008
hba	0.93 (0.67)	0.38 (0.28)	NA (4.95)	0.51	0.87	0.000
ins	0.23 (0.20)	2.10 (1.80)	NA (2.21)	0.79	0.11	0.06
hbb	4.64 (3.38)	4.24 (3.29)	7.38 (1.63)	0.005	0.003	0.07
Mitochondrial loci						
atp6	0.10 (0.01)	1.66 (1.24)	73.75 (5.46)	0.98	0.23	0.000
cytb	6.10 (4.17)	4.86 (3.81)	NA (3.14)	0.007	0.02	0.02
cytox1	2.00 (1.75)	1.71 (1.60)	NA (1.02)	0.14	0.17	0.26
cytox2	14.26 (9.81)	15.86 (13.74)	NA (0.35)	0.000	0.000	0.64
cytox3	1.00 (0.72)	0.16 (0.14)	NA (2.82)	0.43	0.87	0.03

The estimation of $\hat{r}$ uses the correction formula of Lanave *et al.* (1984). NA indicates that the Lanave technique fails because of the incompatibility of the data with the assumptions of the method. The numbers in parentheses are estimates of $\hat{r}$ that do not use a correction formula. The bootstrap results are the probability of observing the data under the assumption of taxonomic exchangeability.

The results of the calculations of $\hat{r}$ are presented in Table 4. For the first position the range of estimated values of R is $0.10 < \hat{r} < 14.26$. For the second position the range is $0.16 < \hat{r} < 15.87$. At this level, cytochrome oxidase 2 and prolactin predictably exhibit high values of $\hat{r}$, as would be expected given the results of the protein analysis. The effects of the correction formula may be seen by comparing these estimates of R with those calculated without using any correction. As might be expected, the correction formula generally inflates the estimated value of R, sometimes by a considerable amount. Therefore, correction formulas themselves may be causing us to underestimate the values of R.

For the third position the fraction of sites that are different is sufficiently large that correction formulas should not be used. A direct examination of the values of R estimated without the correction formula does provide some interesting observations. Notice, initially, that the third position of cytochrome oxidase 2 does not exhibit a particularly high value of $\hat{r}$, unlike the first two positions and the protein. However, for some other loci the opposite is true. For example, α-hemoglobin has one of the lowest estimated values of R for the protein, whereas the third position exhibits one of the higher values of $\hat{r}$.

Since the correction formulas cannot be trusted, we require a method for testing for heterogeneity of rates that does not depend on our ability to correct for multiple substitutions. A test that appears to accomplish this end is based on a resampling scheme similar to those used in bootstrap techniques (Efron 1982). The test makes use of the fact that the sequences from species that are evolving at the same rate should exhibit "taxonomic exchangeability." Suppose we have four species in a particular star phylogeny. Suppose we observe the bases (A,A,A,T) in the four species at a number of sites. (The four bases in this vector represent the bases seen at a particular site in species 1, 2, 3 and 4, in that order.) This should occur with about the same frequency as we observe the bases (A,A,T,A), (A,T,A,A) and (T,A,A,A). We can generate sequences that exhibit taxonomic exchangeability on the computer by choosing sites at random from the sequences of a group of species from a star phylogeny and then randomly permuting the bases. After repeating this process enough times to generate a set of sequences of the same length as the original sequences, the value of $\hat{r}$ can be calculated for these randomly permuted sequences (without, of course, applying a correction formula). If this process is repeated a large number of times, we end up with an empirical distribution of $\hat{r}$ against which we can compare the value of $\hat{r}$ calculated from the data. It is an easy matter to record the fraction of times that the randomly generated $\hat{r}$ exceeds the value observed in the data. This provides the significance of the observed value of $\hat{r}$.

Table 4 gives the significance values for the uncorrected values of $\hat{r}$ for the three positions. If we arbitrarily decide to say that the value of $\hat{r}$ is significantly different from the expected value if the observed value of $\hat{r}$ is exceeded in 1% of the simulation cases, then Table 4 indicates that three of nine of the $\hat{r}$ values are significant for the first, second and third positions. A remarkable aspect of this result is the similarity of the three positions. Under some mechanistic models of evolution, for example the neutral allele theory, one might have expected the third position to exhibit more homogeneity of rates than the first two positions.

IMPLICATIONS FOR THE MECHANISM OF MOLECULAR EVOLUTION

The most intriguing aspects of these results come from making various comparisons: mitochondrial *vs.* nuclear, proteins *vs.* nucleotides. What will emerge is a picture of molecular evolution that differs in significant ways from the usual interpretations based on the neutral allele theory.

In comparing the mitochondrial to nuclear loci several things stand out. The first concerns rates of evolution. As has been amply documented previously (e.g., see BROWN *et al.* 1982), silent substitution rates in the mitochondria exceed those in the nucleus. Because of the problems with correction formulas, is is difficult to quantify the extent of the difference using our data. A good index of the difference is the fraction of third position sites that differ. For our data, the average fraction of third position sites that differ in the mitochondria is 0.505, and for nuclear genes it is 0.317. If we could change these fractions into numbers of substitutions per site, the difference would undoubtedly be more dramatic. However, and this is the main point, the protein data does not exhibit a similar difference in substitution rates between the mitochondria and nuclear loci. For example, the average fraction of amino acids that differ between the nuclear loci is 0.240, and for mitochondrial loci it is 0.190. This comparison is not totally satisfactory because the proteins used in these comparisons are different, and different proteins generally evolve at different rates. However, a possible result, that the mitochondrial proteins evolve at a much higher rate than the nuclear proteins is not seen in this data.

The mitochondrial and nuclear proteins also do not seem to exhibit any striking differences in the evolutionary properties that are examined in this study. For example, in the mitochondria, 11 of 15 comparisons yielded significant interactions in the substitutions in the first and second positions in the codon analysis (Table 3), whereas for the nuclear genes the figure is 10 of 18. For the mitochondrial loci, three of five of the loci have values of $\hat{r}$ for proteins that are significantly larger than one, whereas three of four of the nuclear loci have significant $\hat{r}$ values (Table 2). Thus, the dynamics of amino acid substitutions appear to be remarkably similar for nuclear and mitochondrial loci. Both exhibit significant rate variation between lineages, and both show excesses of double substitutions. Thus, we cannot reject the hypothesis that the forces responsible for nuclear and mitochondrial protein evolution are similar.

The nucleotide comparisons show a rate difference between the mitochondrial and nuclear loci at the third position, but otherwise the two groups of loci are remarkably similar in their dynamics. For example, the bootstrap that tests for heterogeneity of rates for the third position found two of four significant departures for nuclear loci and two of five for the mitochondrial loci (Table 4). The behavior of δ_{bf} for the mitochondrial and nuclear loci appears to be similar. As with the protein data, the similarity in the mitochondrial and nuclear dynamics suggests that similar forces might be operating, although the average rate in the mitochondria is obviously higher.

These patterns suggest the following interpretation. The fact that the mitochondrial proteins appear to evolve at a rate similar to nuclear proteins while the mitochondrial silent sites evolve at a higher rate suggests that the evolution of proteins and silent sites are somewhat uncoupled. For whatever reason the mitochondrial silent sites are evolving more rapidly, it does not lead to a speedup of the evolution of proteins. The argument that replacement and silent substitutions are uncoupled is further supported by the fact that the high values of $\hat{r}$ are uncorrelated for silent and replacement substitutions. For example, α-hemoglobin gives a value of $\hat{r}$ of 0.94 for replacement substitutions and 19.42 for silent substitutions (Table 3). Conversely, cytochrome oxidase 2 gives an $\hat{r}$ of 35.14 for replacement substitutions and 1.22 for silent substitutions. A similar lack of correlation is seen between positions two and three in Table 4.

This apparent uncoupling of the rates of silent and replacement substitutions presents a dilemma for the neutralist explanation. Variation in parameters, such as population size or generation time, that might lead to variation in the rates of evolution under neutrality should affect all loci and all sites within a locus equally. The fact that different loci and different sites within a locus exhibit different values of $\hat{r}$ would appear to rule out neutrality with fluctuating population sizes or generation times as a viable model. Fluctuations in the mutation rate from locus to locus could account for the fact that $\hat{r}$ varies across loci, but cannot account for the fact that the values of $\hat{r}$ for silent and substitution sites within a locus are uncorrelated. A related problem is the fact that the silent substitutions do not exhibit a lower variance in rates as measured by $\hat{r}$ or by the bootstrap when third positions are compared to second positions

(Table 4). All of these observations appear to argue against the neutral model as it is presently defined (KIMURA 1983).

In three recent papers (GILLESPIE 1984a,b, 1986a,b) I have argued that natural selection would be a likely candidate for protein evolution since models of protein evolution by natural selection, unlike neutral allele models, predict the high values of R typically seen in the data. Some of the values of R estimated in this paper are considerably higher than any seen before, putting the neutral theory in even greater jeopardy.

If we were persuaded by the heterogeneities in the data to reject the neutral model, there are some phenomena that must be accounted for by purely selectionist arguments. One phenomena that need not be accounted for is the higher rate of silent substitutions over amino acid substitutions. Presumably, the evolution of proteins is primarily involved with the behaviors of the protein molecule itself. The substitutions of bases that do not change amino acids could function to alter codon usage, the secondary structure of the message or the DNA, the transcription rate, the avoidance of CG neighbors, or a host of other properties of the DNA. There is no way to say, *a priori*, which of these two levels of evolution should proceed more rapidly. However, there is an obvious reason why mitochondrial silent substitutions should occur more frequently: the nuclear DNA is intimately bound to histones that are themselves extraordinarily conservative, whereas the mitochondrial DNA is naked. Thus, the nuclear DNA must be more constrained, causing fewer advantageous mutations to be available as candidates to become fixed in the population, hence the lower rate of evolution.

The unexpected behavior of δ_{bf} deserves some additional comment. The fact that δ_{bf} increases dramatically with an increasing number of nucleotide differences suggests that nucleotide evolution always involves changes in base frequencies. That is, base frequencies are not conserved quantities in evolution as assumed in most of the models of molecular evolution. The conservation of base frequencies is a property of mutation-driven models, such as the neutral allele theory. Unfortunately, little is known about the behavior of base frequencies under models of natural selection.

This paper has benefited from the comments of many individuals. Among them are MASARU IIZUKA, CHUCK LANGLEY, HIROTSUGA MATSUDA, TOMOKO OHTA, NAOYUKI TAKAHATA, MICHAEL TURELLI and, indirectly, MOTOO KIMURA.

LITERATURE CITED

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.

EFRON B., 1982 The jackknife, the bootstrap and other resampling plans, Vol. 38. CBMS-NSF Regional Conference Series on Applied Mathemathics. SIAM, Philadelphia.

FITCH, W. M., 1971 Rate of change of concomitantly variable codons. J. Mol. Evol. **1:** 84–96.

FITCH, W. M. and E. MARGOLIASH, 1966 The construction of phylogenetic trees. II. How well do they reflect past history? Brookhaven Symp. Biol. **21:** 217.

GILLESPIE, J. H., 1984a Molecular evolution over the molecular landscape. Evolution **38:** 1116–1129.

GILLESPIE, J. H., 1984b The molecular clock may be an episodic clock. Proc. Natl. Acad. Sci. USA **81:** 8009–8013.

GILLESPIE, J. H., 1986a Statistical aspects of the molecular clock. pp. 255–272. In: *Evolutionary Processes and Theory*, Edited by S. KARLIN and E. NEVO. Academic Press, New York.

GILLESPIE, J. H., 1986b Natural selection and the molecular clock. Mol. Biol. Evol. **3:** 138–155.

GILLESPIE, J. H. and C. H. LANGLEY, 1979 Are evolutionary rates really variable? J. Mol. Evol. **13:** 27–34.

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

KAPLAN, N., 1983 Statistical analysis of restriction enzyme map data and nucleotide sequence data. pp. 75–106. In: *Statistical Analysis of DNA Sequence Data*, Edited by B. S. WEIR. Marcel Dekker, New York.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. pp. 1–367. Cambridge University Press, Cambridge.

LANAVE, C., G. PREPARATA, C. SACCONE and G. SERIO, 1984 A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20:** 86–93.

LANGLEY, C. H. and W. M. FITCH, 1974 An estimation of the constancy of the rate of molecular evolution. J. Mol. Evol. **3:** 161–177.

MIYATA, T., 1982 Evolutionary changes and functional constraints in DNA sequences. pp. 233–266. In: *Molecular Evolution, Protein Polymorphism and the Neutral Allele Theory*, Edited by M. KIMURA. Springer-Verlag, Berlin.

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.

MIYATA, T. and T. YASUNAGA, 1980 Molecular evolution of mRNA: a method of estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16:** 23–36.

OHTA, T. and M. KIMURA, 1971 On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. **1:** 18–25.

TAVARE, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences. In: *Lectures on Mathematics in the Life Sciences*, Vol. 17, Edited by R. M. MIUR. American Mathematical Society, Providence, Rhode Island. In press.

WATTERSON, G. A., 1982a Substitution times for mutant nucleotides. J. Appl. Probab. **19A:** 59–70.

WATTERSON, G. A., 1982b Mutant substitutions at linked nucleotide sites. Adv. Appl. Prob. **14:** 206–224.

WILSON, A. C., S. S. CARLSON and T. J. WHITE, 1977 Biochemical evolution. Annu. Rev. Biochem. **46:** 573–639.

ZUCKERKANDL, E. and L. PAULING, 1965 Evolutionary divergence and convergence in proteins. pp. 97–166. In: *Evolving Genes and Proteins*, Edited by V. BRYSON and H. J. VOGEL. Academic Press, New York.

Communicating editor: M. T. CLEGG