J. molec. Evolution 2, 123—136 (1973)

Is the Fixation of Observable Mutations Distributed Randomly among the Three Nucleotide Positions of the Codon?

Walter M. Fitch

Department of Physiological Chemistry, University of Wisconsin, Madison, Wisconsin, U.S.A.

Received April 3, 1973

Summary. The distribution among the three nucleotide positions of the codons of 642 mutations fixed during the descent of 49 sequences of cytochrome *c* was examined. This was compared to the distribution expected if the number of ways of getting a selectively acceptable amino acid alternative from a single nucleotide replacement at each coding position were random, *i.e.* proportional to the total number of ways of changing the encoded amino acid by a single nucleotide replacement at each coding position. It was found that the observed distribution was significantly different from random, there being 40% more mutations in the first coding position than in the second whereas one would have expected 10% more in the second than in the first. The probability of the result occurring by chance is $<10^{-6}$.

The same test was made on the distribution of 347 mutations fixed in the descent of 19 sequences of alpha hemoglobin and 286 mutations fixed in the descent of 16 beta and 4 delta hemoglobins. The result for the alpha hemoglobins was a similar non-randomness but the probability of its occurring by chance rose to 0.005. The result for the beta-delta hemoglobins was in the same direction but was not significant ($p = 0.3$). The degree of non-randomness among the three genes in the distribution of fixations over the three nucleotide positions of their codons appears to be correlated (negatively) with their rates of evolution, the plasticity required of the molecule to adapt to new environments, and the recency of exploitation of opportunities for change in functional specificity provided by such processes as gene duplication.

Key words: Nucleotide Substitutions — Evolution — Codons — Cytochromes *c* — Hemoglobins — Fibrinopeptides A.

Previous studies (Fitch and Markowitz, 1970; Fitch and Margoliash, 1967) have shown that the distribution of observable mutations among the codons of a gene is non-random (i.e. not a simple Poisson process) but rather, that some codons are invariable and that among the remainder, the codons do not all have the same degree of variability. This has been shown for cytochrome *c* (Fitch and Markowitz, 1970), fibrinopeptide A (Margoliash and Fitch, 1970), α-hemoglobin (Fitch, 1972a) and β-hemoglobin (Fitch, 1972b). At the other extreme, there has been shown a non-randomness in the character (type) of nucleotide replacements (Fitch, 1967; Vogel, 1969) there being a much elevated frequency of G→A replacements

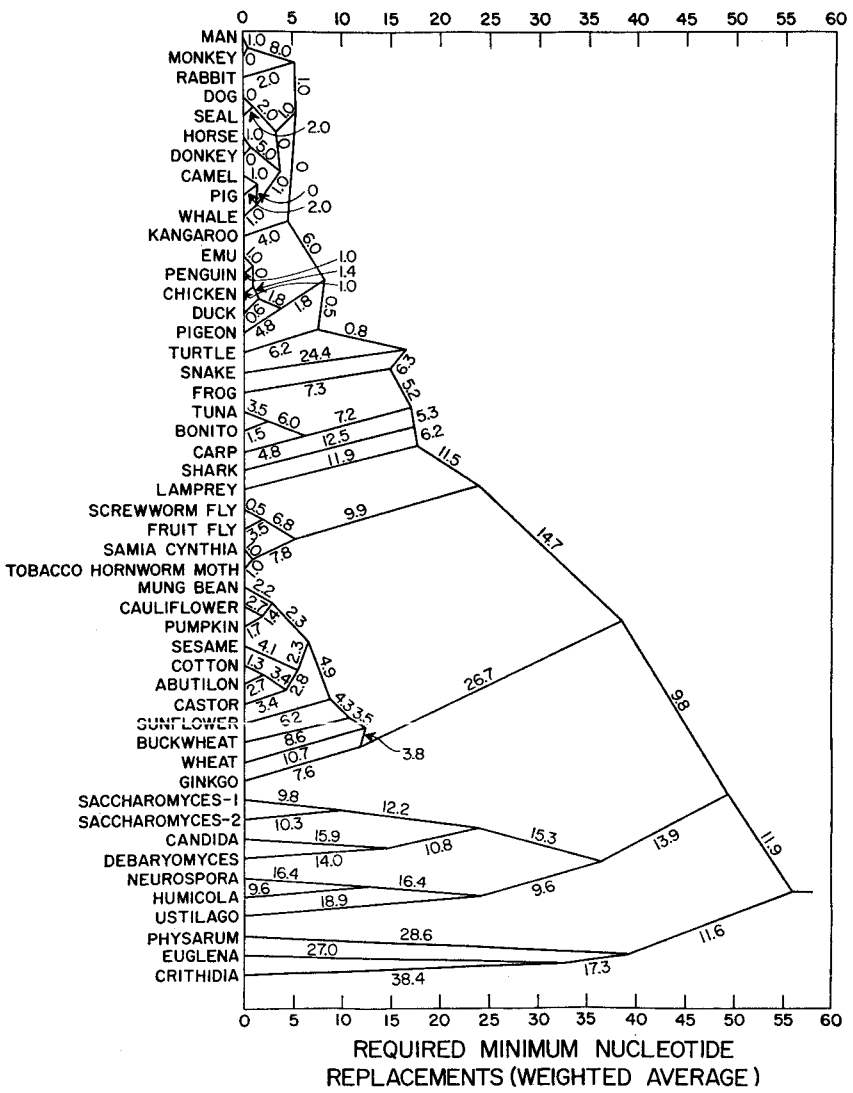
compared to that expected if each nucleotide type were equally subject to replacement and if each of the three replacing types were equally acceptable on the average. It has since been shown (Fitch, 1972a) that the distribution of nucleotide replacement type in both the first and second nucleotide positions are, however, broadly consistent with that expected if we assume that the nucleotide composition of the gene has approached an equilibrium condition. In the intermediate range, one may ask if the observable mutations are fixed in the three nucleotide positions of the codon in a random manner. I shall show that they are not, there being a greater than random fixation of mutations in the first nucleotide position and a smaller than random in the second, the effect being most marked in cytochrome *c*, intermediate in alpha hemoglobin and not detectable in beta hemoglobin.

Methods

To determine the expected rate of substitution in the various nucleotide positions of the codon, one needs to know the codon composition of the gene fixing the mutation and the number of ways each codon can change. The substitutions being examined are those fixed in the evolution of the 49 cytochromes *c* shown in Fig. 1. For the purposes of this work, then, the most appropriate codon composition short of re-examining every ancestral sequence after each fixation, is to use the codon composition of the ultimate ancestral gene. This sequence is shown in Scheme 1 and its composition is shown in Table 1. The ancestral sequence was determined by the method of Fitch (1971) which Hartigan (1973) has provided a mathematical proof of.

A few caveats are in order. First of all, the observational data are counts by codon position of the nucleotide replacements necessary to account for the observed amino

Fig. 1. The phylogeny of the 49 species as it was assumed. The minimum number of mutations required to account for the descent from their ancestral cytochromes *c* is shown along each leg. Their sum is 642. The numbers are not integers due to distributing mutations over several legs where more than one possible most parsimonious solution exists. The sum of the numbers on the legs connecting two species then becomes the minimum phyletic distance. This may be considerably more than the minimum pairwise distance which takes no account of the phylogeny. For example, the pig-*Euglena* minimum phyletic and pairwise distances are 138.1 and 59 nucleotide replacements respectively. The following are the cytochromes *c*, in the order in which they appear in the figure: man, Matsubara and Smith (1963); *Macaca mulatta*, Rothfus and Smith (1965); rabbit, Needleman and Margoliash (1966); dog, McDowell and Smith (1965); *Mirounga leonina*, Augusteyn *et al.* (1972); horse, Margoliash *et al.* (1961); donkey, Walasek and Margoliash unpublished; *Camelus dromedarius*, Sokolovsky and Moldovan (1972); pig, Stewart and Margoliash (1965); *Rachianectes glaucus*, Goldstone and Smith (1966); *Macropus canguru*, Nolan and Margoliash (1966); *Dromaeus novaehollandia*, Augusteyn (1973); *Aptenodytes patagonica*, Chan, Tulloss and Margoliash, unpublished; chicken, Chan and Margoliash (1966a); *Anas platyrhyncos* and pigeon, Chan, Tulloss, and Margoliash, unpublished; *Chelydra serpentina*, Chan *et al.* (1966); *Crotalus adamanteus*, Bahl and Smith (1965); frog, Chan, Walasek, Barlow, and Margoliash, unpublished; tuna, Kreil (1963); *Katsuwonus pelamis*, Nakayama *et al.* (1972); *Cyprinus carpio*, Gurther and Horstman (1970); *Squalus sucklii*, Goldstone and Smith (1967); *Entosphenus tridentatus*, Nolan, Fitch, Uzzell, Weiss, and Margoliash, submitted for publication; *Haematobia irritans*, Chan, Tulloss, and Margoliash, unpublished; *Droso-*



phila melanogaster, Nolan, Weiss, Adams, and Margoliash, unpublished; *Samia cynthia*, Chan and Margoliash (1966b); *Protogarce sexta*, Chan (1970); *Phaseolus aureus*, Thompson *et al.* (1970a); *Brassica oleracea*, Thomson *et al.* (1971b); *Cucurbita maxima*, Thompson *et al.* (1971a); *Sesamum indicum*, Thompson *et al.* (1970b); *Gossypium barbadense* and *Abutilon theophrasti*, Thompson *et al.* (1971c); *Ricinus communis*, Thompson *et al.* (1970b); *Helianthus annus*, Ramshaw *et al.* (1970); *Fagopyrum esculentum*, Thompson *et al.* (1971b); wheat, Stevens *et al.* (1967); *Ginkgo biloba*, Ramshaw *et al.* (1971); *Saccharomyces cerevisiae/oviformis iso*-I, Narita and Titani (1969); Lederer *et al.* (1972); ibid, *iso*-II, Stewart, Putterman, and Margoliash, unpublished; *Candida kruzei*, Narita and Titani (1968); *Debaryomyces kloeckeri*, Sugeno *et al.* (1971); *Neurospora crassa*, Heller and Smith (1966); *Humicola lanuginosa*, Morgan *et al.* (1972); *Ustilago sphaerogena*, Bitar *et al.* (1972); *Physarum polycephalum*, Lin, Niece, and Fitch, unpublished; *Euglena gracilis*, Pettigrew (1973), Lin, Niece, and Fitch (1973); *Crithidia oncopelti*, Pettigrew (1972)

Scheme 1. Ancestral eukaryotic cytochrome *c* gene. The sequence was determined by the method of Fitch (1971) using the phylogeny shown in Fig. 1. Nucleotides are represented as follows: A = adenine; B = not adenine = C/G/U; C = cytosine; D = not cytosine = A/G/U; G = guanine; H = not guanine = A/C/U; U = uracil; V = not uracil = A/C/G; R = purine = A/G; Y = pyrimidine = C/U; S = A/C; T = A/U; W = U/G; Z = C/G. The ten codons whose ambiguity contains codons for more than two amino acids are as follows: −8 = Lys/Ala/Thr/Glu (only Lys and Ala known in this position); −5 = Glu/Asp/Ala; −4 = Pro/Gly/Ser/Arg/Ala/Trp (only Pro, Gly, Ser known in this position); 42 = Thr/Pro/Ser; 50 = Asn/Asp/His; 35 = Lys/Ala/Thr/Glu (only Lys and Ala known in this position); 85 = Ile/Leu/Phe; 92 = Lys/Glu/Asn/Asp (Asp unknown in this position); 100 = Lys/Asn/Thr; 103 = Lys/Asn/Glu/Asp/Tyr (Tyr unknown in this position). Underlined residues are those which are unvaried among the eukaryotic cytochromes *c*. Positions 14, 27, 67, 75, and 87, not previously known to have varied, have been shown to be varied among some protozoa, *Crithidia* (Pettigrew, 1972) and *Euglena* (Pettigrew, 1973; Lin, Niece, and Fitch, 1973). Position 37 appears to be lysine in *Physarum* (Lin, Niece, and Fitch, unpublished).

	−9								−1
	ZCX Pro Ala	RSR −8	RCR Thr Ala	YGY Arg Cys	GSX −5	BZR −4	UUY Phe	ZCR Pro Ala	CCX Pro
GGX Gly	GAY Asp	GCA Ala	RAR Lys Glu	AAR Lys	GGX Gly	GAR Glu	AAR Lys	CUH Leu	10 UUY Phe
AAR Lys	RCG Thr Ala	AGR Arg	UGY Cys	GCR Ala	CAR Gln	UGY Cys	CAY His	ACY Thr	20 GYA Ala Val
GAR Glu	AAR Lys	GGY Gly	GGR Gly	ZCR Pro Ala	SAY Asn His	AAR Lys	GUA Val	GGX Gly	30 CCX Pro
AAY Asn	CUX Leu	YAY His Tyr	GGX Gly	SUY Ile Leu	UUY Phe	GGY Gly	CGX Arg	AAR Lys	40 UCX Ser
GGX Gly	HCR 42	GYX Ala Val	CCR Pro	GGX Gly	UAY Tyr	WCX Ala Ser	UAY Tyr	ACX Thr	50 VAY 50
GCX Ala	AAY Asn	AAA Lys	AAY Asn	RSG 55	GGY Gly	AUA Ile	ACA Thr	UGG Trp	60 GAR Glu
GAR Glu	GAR Glu	ACY Thr	CUB Leu	UUY Phe	GAA Glu	UAY Tyr	CUX Leu	GAR Glu	70 AAY Asn
CCX Pro	AAR Lys	AAR Lys	UAY Tyr	AUH Ile	CCX Pro	GGR Gly	ACX Thr	AAR Lys	80 AUG Met
GCH Ala	UUY Phe	GCX Ala	GGX Gly	HUY 85	AAR Lys	ASR Lys Thr	GCR Ala	AAR Lys	90 GAR Glu
CGX Arg	RAX 92	GAY Asp	CUH Leu	AUH Ile	GCX Ala	UAY Tyr	CUG Leu	AAR Lys	100 ASX 100
GYR Ala Val	ACX Thr	DAX 103	GAR Glu						

Table 1. Frequency of codons in ancestral cytochrome *c*

First	Second								Third
	A		G		C		U		
A	Lys	7.98	Arg	0.50	Thr	2.67	Ile	1.67	A
		7.22		0.50		2.42	Met	1.00	G
	Asn	2.77	Ser	—		1.88	Ile	1.08	C
		2.77		—		1.88		1.08	U
G	Glu	5.72	Gly	2.59	Ala	5.30	Val	1.88	A
		4.98		2.59		4.22		0.38	G
	Asp	1.52		3.50		1.58		0.12	C
		1.52		3.50		1.58		0.12	U
C	Gln	0.50	Arg	0.59	Pro	2.38	Leu	1.17	A
		0.50		0.59		2.38		1.83	G
	His	1.17		0.75		1.12		1.92	C
		1.17		0.75		1.12		1.92	U
U	End	—	End	—	Ser	0.63	Leu	—	A
		—	Trp	1.09		0.63		—	G
	Tyr	2.85	Cys	1.25		0.38	Phe	2.67	C
		2.85		1.25		0.38		2.67	U

acid differences. As such, they are minimal estimates of the total number fixed. Our analysis involves relative comparisons among the three codon positions. Therefore a minimal estimate is no problem in the absence of any bias that would make the fraction of the total of the amino acid changing replacements counted differ among the nucleotide positions.

Second, wherever there is ambiguity in the nature of the ancestral codon, the assignment is distributed equally among all possible codons (excluding terminators if there had been any) consistent with the ambiguity. If then, one codon for an amino acid were preferentially used over another throughout the sequence and could give rise to codons for other amino acids in a different number of ways, that could introduce a small bias. For example GGG glycine has one less first nucleotide alternative than other glycine codons because UGG is a terminator. The only amino acids for which this is even potentially a problem are arginine, glycine, leucine and serine.

A similar (third) caveat would apply to cases such as the ambiguous histidine-tyrosine choice at ancestral cytochrome-33 where there are no third nucleotide options open to the tyrosine codon since they would lead to terminators. However, this particular situation does not repeat itself elsewhere in the sequence and the total number of such codons where an unknown bias could operate, if there is any bias at all, represents only about 10% of the total positions. Moreover, even if it existed, the bias, to be important, would have to work in the same direction among different pairs of amino acids.

Fourth, in those cases involving amino acids with six-fold degeneracy of coding (arginine, leucine and serine) the procedure requires an *a priori* choice as the procedure could count too few fixations if an extant codon were given ambiguity in more than one nucleotide position destined to mutate. However, there may be no information at a particular position to determine how one should restrict the encoding. For example, Arg-38, Leu-68, and Arg-91 are unvaried over all cytochrome species examined. In such cases, the four-fold ambiguous coding is automatically assigned i.e. CGX for Arg, CUX for Leu, and UCX for Ser. Clearly this will introduce a small bias into the composition of the ancestral codons in the following sense. The Arginine codons AGA and AGG (AGR) can mutate to other amino acids in six and seven ways

Table 2. Ways in which a single nucleotide replacement can change the amino acid encoded

Amino acids (codons)	Pattern	No. of codons	First	Second	Third	Total
Met	3/3/3	1	3	3	3	9
Asn, Asp, His, Phe, Ser[1]	3/3/2	10	30	30	20	80
Cys, Ile	3/3/1	5	15	15	5	35
Ala, Arg[2], Gly[3], Leu[4], Pro, Ser[5], Thr, Tyr, Val	3/3/0	27	81	81	—	162
Trp	3/2/2	1	3	2	2	7
Ser (UCG)	3/2/0	1	3	2	—	5
Ser (UCA)	3/1/0	1	3	1	—	4
Arg[6], Gln, Glu, Lys	2/3/2	7	14	21	14	49
Arg[7], Leu[8], Gly[9]	2/3/0	4	8	12	—	20
Leu (UUG)	2/2/2	1	2	2	2	6
Leu (UUA)	2/1/2	1	2	1	2	5
Arg (AGA)	1/3/2	1	1	3	2	6
Arg (CGA)	1/3/0	1	1	3	—	4
Total		61	166	176	50	392

[1] AGY; [2] CGY; [3] excluding GGA; [4] CUY; [5] UCY; [6] AGG; [7] CGG; [8] CUR; [9] GGA. The letters Y and R imply pyrimidine and purine respectively. The pattern $(i/j/k)$ indicates the number of nucleotides, i, j or k, that can replace the existing nucleotide in the first, second or third coding positions respectively so as to change the amino acid encoded.

The relative probabilities of an observable fixation in the first, second and third nucleotide positions, assuming all codons equally frequent and all amino acid changes equally likely to be fixed, are 0.4235, 0.4490 and 0.1276 respectively. For the complete set of 401 ancestral codons examined (113 cytochrome *c*, 142 α hemoglobin, 146 beta-delta hemoglobins) the probabilities are 0.4180, 0.4520, and 0.1300. The total number of ways those 401 codons could mutate with a change of the encoded amino acid is 2617.77 for an average 6.528 ways/codon. If all codons were equiprobable, there would be $392/61 = 6.426$ ways/codon.

respectively whereas, for the CGX set, CGC and CGU can do so six ways, but CGG and CGA can do so in only 5 and 4 ways respectively. While the actual number of mutations found in the various coding positions would not be affected by this choice, the number expected would be slightly altered. This might be expected to affect the third coding position most (by underestimating their expected number) since the amino acid encoded by CGX could not be altered by a third position change whereas that encoded by AGR could be. To the extent this bias exists, it may be partially overcome by removing unvaried codons when estimating the expected distribution of the observed replacements.

The number of ways that each codon can change in each nucleotide position and change the amino acid encoded is shown in Table 2. Of the $61 \times 9 = 549$ ways that the non-terminator codons can mutate, 392 of those ways change the amino acid encoded, 132 are silent and 25 are to terminators. Of the 392 mutations that produce a change in the encoded amino acid, 166 are in the first position, 176 in the second, and only 50 are in the third. This partition, divided by 392, would give the relative probabilities

Table 3. Distribution of observed fixations by position of nucleotide in codons of cytochrome *c*

	First	Second	Third	Total
I Ways of changing the codons	303.2	336.0	99.8	739.0
II Relative probability	0.4103	0.4547	0.1350	1.000
III Expected distribution (II X 642)	263.4	291.9	86.7	642
IV Found distribution	329.0	234.0	79.0	642
V Chi square ($df = 2$)	16.3	11.5	0.7	28.5
I′ Ways of changing the codons	232.2	258.0	80.8	571.0
II′ Relative probability	0.4067	0.4518	0.1415	1.000
III′ Expected distribution (II′ X 642)	261.1	290.1	90.8	642.0
IV′ Found distribution	329.0	234.0	79.0	642.0
V′ Chi square ($df = 2$)	17.7	10.8	1.5	30.0

Data are for cytochrome *c* using the complete ancestral sequence shown in Fig. 2 as the basis for computing the values in row I. The procedure is explained in the text. The values in the lower half of the table (shown with primes) are similar except that the values in row I′ were computed using the ancestral sequence minus the 26 unvaried codons underlined in Fig. 2. The probability that, by chance, the fixations would distribute themselves as found under the assumptions made is $< 10^{-6}$ in both cases.

of a fixation in the three nucleotide positions (cf. Table 2) assuming that all codons were equally frequent and that all resultant changes were equally likely to be fixed. We can correct for the inaccuracy of the first assumption. The columns in Table 2 labelled First, Second and Third were obtained by multiplying the respective digits of the Pattern column by one for each different codon of that pattern. If, instead, we multiply the digits of the pattern by the number codons of that pattern in the ancestral cytochrome *c* gene as given in Table 1, we get the expected distribution shown in the top row of Table 3. Thus the total number of ways of altering the ancestral amino acid sequence by a single nucleotide change is 739. If the total of 739 is divided into the number of ways of making an amino acid change in each nucleotide position, one obtains the probability (second row of Table 3) of an amino acid changing mutation in that position. The evolution shown in Fig. 1 involves a total of 642 fixations. If that is multiplied by the appropriate probability, one obtains an expected distribution of those fixations as shown in the third row. This is the distribution of fixations to be expected if all alternatives are equally acceptable or, less stringently, if acceptable alternatives are distributed in accordance with those probabilities.

Results and Discussion

The distribution of those mutations as actually found is shown in the fourth row of Table 3. The chi-square of 28.5 for 2 degrees of freedom shows that the found distribution would have a probability of occurring by chance under these assumptions of $< 10^{-6}$. Therefore we can reject the assumption that the mutations fixed are distributed randomly over the three nucleotide positions. Examining these data more closely, we see that the fixations in the third positions are approximately in the number expected so that the large chi-square value is associated with discrepancies in the first two

positions. There we see that whereas we expect more mutations to be fixed in the second nucleotide position than in the first (there being greater opportunities for such changes in the second), in fact there are considerably more mutations fixed in the first than in the second nucleotide position. This appears to have a reasonable biological basis. Mutations in the first position include most of those amino acid changes that are considered as closely related including the hydrophobic residues with a second position U, methionine, valine, leucine, isoleucine and phenylalanine, the residues with a second position C, alanine, serine, threonine and proline, and the hydrophilic residues with a second position A, glutamate, aspartate, glutamine, asparagine, histidine and lysine. The non-randomness is thus viewable as the effect of selection restricting the nature of allowable amino acid replacements at the various sites of a functioning molecule, a conclusion consistent with all the other molecular evolutionary observations discussed in the introduction. King (1972) has pointed out that, according to Sneath's measures of chemical similarity (1966), first position replacements are somewhat more conservative than second position replacements.

The potential bias in the codon assignments where there is insufficient information to restrict the choice on a parsimony principle was alluded to above. To test the extent of that bias, the preceding calculations were repeated using the ancestral cytochrome *c* minus the 26 unvaried positions underlined in Scheme 1. The result, given in the lower half of Table 3, shows that the arbitrary nature of codon assignment to the unvaried codons has almost no effect on the expected distribution of fixations among the three nucleotide positions. If anything, the non-randomness is increased. It should be remarked that this test of bias does not cover every contingency relating to codon assignment for the six-fold degenerate cases. The possible bias in calculating the expected distribution of nucleotide replacements was discussed earlier under the fourth caveat. There is one amino acid, serine, where the choice affects the nucleotide position to which the replacement is assigned. For example, codon 40 codes only for threonine (ACX, 30 taxa) and serine (19 taxa). Now it is immaterial whether the serine is coded AGY or UCX so far as the phyletic location and the minimum number of fixations (two, as it turns out) are concerned. However, had serine been assigned the codons AGY instead of UCX, those two fixations would have been recorded in the second nucleotide position rather than in the first. On the other hand, an ancestral AGY assignment would also have led to a small, partially compensatory increase in the expectation for second position fixations. All in all, since there are few cases of the type not allowed for, and since the greatest effect of any such bias would be on the third nucleotide position where it should lead to an underestimation of the number of third position fixations but doesn't, and whereas the probability of the result is already very low, I regard the result as clearly indicating that the observable fixa-

Table 4. Distribution of observed fixations by position of nucleotide in codons of alpha hemoglobin

	First	Second	Third	Total
I Ways of changing the codons	396.6	419.4	114.1	930.1
II Relative probability	0.4264	0.4509	0.1226	1.000
III Expected distribution (II X 347)	148.0	156.5	42.6	347.0
IV Found distribution	177.0	138.0	32.0	347.0
V Chi square ($df = 2$)	5.7	2.2	2.6	10.5

Data are for 19 alpha hemoglobins using the ancestor common to carp and higher vertebrates for computing the values in row I. This is ancestor B in Fig. 1 of Fitch (1972a) which gives the phylogeny and the references to the sequences used. The probability that, by chance, the fixations would distribute themselves as found under the assumptions made is $\sim 5 \times 10^{-3}$. Removal of the unvaried codons would increase p to ~ 0.01.

Table 5. Distribution of observed fixations by position of nucleotide in codons of beta-delta hemoglobin

	First	Second	Third	Total
I Ways of changing the codons	399.7	433.6	128.3	961.5
II Relative probability	0.4157	0.4509	0.1334	1.000
III Expected distribution (II X 286)	118.9	129.9	38.2	286.0
IV Found distribution	131.0	123.0	32.0	286.0
V Chi square ($df = 2$)	1.2	0.3	1.0	2.5

Data are for 16 beta and 4 delta hemoglobins using the ancestor common to frog and higher vertebrates (ancestral amniote) for computing the values in row I. This is ancestor B in Fig. 4 of Fitch (1972b) which gives the phylogeny and the references to the sequences used. The probability that, by chance, the fixations would distribute themselves as found under the assumptions made is ~ 0.3. Removal of the unvaried codons in calculating the expected distribution would decrease p to ~ 0.2.

tions are not distributed randomly among the three nucleotide positions of codons for cytochrome *c*.

One may ask whether what is true for the cytochrome *c* gene is true for other genes. In Table 4 is shown a comparable calculation for alpha hemoglobin. The result is the same except that one's confidence in the non-randomness is markedly reduced (probability of occurring by chance is up to 0.005). Given the reservations regarding possible bias in some codon assignments, I would think it best to regard the non-randomness conclusion as somewhat tentative.

When the same computation is performed for beta hemoglobins (Table 5) the result is that no significant non-randomness is present although in each position the observations lie on the same side of expectation as for the other two genes.

Table 6. Distribution of observed fixations by position of nucleotide in codons of all genes examined

	First	Second	Third
Expected	529.6	574.0	171.4
Found	637	495	143
Chi square	21.8	10.9	4.7

Data are for the 1275 fixations observed in all genes presented in Tables 3–5 except that expected values are determined only on those codons that have varied, i.e. the unvaried positions have been removed. The probability of the chance occurrence of these results is 4×10^{-6}, 2×10^{-3} and 0.03 for the first, second and third positions respectively. The overall probability for chi-square $= 37.4$ $(df = 2)$ is 5×10^{-10}.

The data for all three observations have been pooled (Table 6) for the cases where expectations are calculated with the unvaried positions removed from consideration. Since the observations for any one nucleotide position all lie on the same side of expected, it is not surprising that the result is even more significant. Indeed, the observations are sufficiently numerous that even the third position changes can be seen to be significantly ($p = 0.03$) underrepresented compared to expectation. Moreover, the one bias the direction of whose effect is known, is the probable slight overutilization of four-fold degenerate codons relative to the two-fold degenerate codons for arginine, leucine and serine. This means that our expected number of third position replacements is itself, if anything, underestimated, making $p = 0.03$ a conservative estimate, i.e. larger than it is in fact. It should be notet that overall, the found overall frequency of nucleotide replacements is 0.4996, 0.3882, and 0.1122 for the first, second and third nucleotide positions respectively. These values are considerably different from those given in Table 2 as expected under the assumption that replacements would be distributed in proportion to the opportunities presented.

The distinctness of the results on the three genes is remarkable. They seem, however, to be plausible when viewed from either of the two following perspectives. In one view, we can assume that with the continuing fixation of mutations over long periods of time, we approach a point where more and more of the possible mutations prove to be malefic i.e., too deleterious to survive. This may result because no further deterioration can be permitted or, perhaps more likely, because one has approached a limit to the perfectibility of the protein given the environment and the starting material. In an alternative view, we can assume that the extent of the historic plasticity of gene may reflect the degree of dissimilarity among the environments to which the gene product has had to adapt.

In explaining the cytochrome *c* results, these two (not necessarily mutually exclusive) considerations suggest (view number 1) that, being an

ancient gene, cytochrome *c* is well perfected and only a narrow range of alternatives, if any is allowed in a given amino acid position. This narrow range is reflected in the preponderance of fixations involving first position mutations (val–leu, ala–thr, asp–asn, etc.). The pertinence of the "if any" is reflected in the limited number of covarions in cytochrome *c* as shown by Fitch and Markowitz (1970). These considerations also suggest (view number 2) that, being encased in the mitochondrion, cytochrome *c* may well have had a more uniform environment and therefore less need to readapt than cytoplasmic proteins, particularly hemoglobin which, by virtue of its localization, comes close to being extracellular.

The basis for the alteration of a gene product's sequence could involve not only a change in the environment, but a change in function or specificity and it could well be that immediately following such a change the gene is more plastic until its new role is better adapted. This process may initially require a gene duplication that permits specialization and the heme-containing globins may illustrate the case in a way that is consistent with the apparent distinctness in plasticity among the gene products examined. Such a view will be briefly outlined.

Consider an ancient, oxygen-carrying, heme-containing globin that, with increasing organismal complexity, found it useful to have one form that worked optimally intracellularly, particularly in occasionally anoxic muscle tissue, and one extracellularly. Thus, *via* a gene duplication, that gene could give rise to myoglobin on the one hand and hemoglobin on the other. Later, the requirements of efficiency of oxygen transport selected for polymeric hemoglobins and cooperativity through tetramers composed of two distinct hemoglobins. Thus, again *via* gene duplication (or heterosis followed by translocation) and shortly after the appearance of the vertebrates, there arose two hemoglobins, one to be called alpha hemoglobin and one that could interact with it. Following still further development of reproductive complexity, it became advantageous that the alpha interacting hemoglobin on the distal side of the placental barrier differ from that on the proximal side, and about the time of the appearance of the mammals, there arose two corresponding loci, which I shall call gamma and beta-delta respectively. The latter has since shown a further duplication into a beta and a delta locus, the advantage of which is obscure. The gamma locus is expressed in the fetal stages of development. If the views are correct, one might predict that the embryonic epsilon hemoglobin, which is expressed in the earliest stages of embryonic development will prove to be more closely related to the gamma than to the other loci.

The crucial point regarding this story is that, prior to the duplication of a locus, its product, designed to serve under differing conditions, was probably a compromise among the forms best designed for the individual conditions. On the other hand, duplication permitted a new plastic phase

to ensue during which more optimal forms were sought. Similar occurrences presumably took place as a primitive non-specific protease gave rise, through duplication, to trypsin, chymotrypsin and other specialist enzymes.

Now I would suggest, although in principle all hemoglobin genes are equally old, and although there may be gene duplications about which we are unaware, and although the degree of plasticity conferred through a gene duplication must be enormously variable, that the peculiarly mammalian development of separate fetal and adult hemoglobin forms could well have invested that beta-delta hemoglobin line with a greater current plasticity than remains in the alpha hemoglobin line, particularly where the latter must interact effectively with all the other hemoglobins while they in turn need interact effectively only with alpha hemoglobin.

If one is willing to consider the preceding view plus the likelihood that the myoglobin-hemoglobin gene duplication was more recent than the advent of eukaryotic cytochrome *c*, then the relative randomness among the three groups in the positions at which nucleotide replacements are observed is just what one would expect. This is only a view consistent with the data, not a conclusion from them but there is other information consistent therewith. The relative rates of evolution (nucleotide replacements/gene/unit time) are, in increasing order, cytochrome *c*, alpha hemoglobin and beta hemoglobin (Barnabas *et al.*, 1971; Fitch, 1972b; Langley and Fitch, 1973) and among the hemoglobins, human variants that survive to time of sampling are significantly less frequent in the alpha than the beta chain (Vogel, 1972). Moreover the hemoglobins have more covarions than cytochrome *c* (Fitch, 1972b). It does appear, at least for the three genes examined, that the extent of non-randomness in the distribution of fixations among the three nucleotide positions of the codon is inversely related to the rate of evolution of the gene. Said another way, the more restricted the selectively acceptable opportunities for change are, the more non-random is the distribution of those mutations among the nucleotide positions of the codon.

Acknowledgements. The technical assistance of Mr. Frank Iltis is gratefully acknowledged. This work received support from NSF grant GB 32274X.

References

Augusteyn, R. C.: Biochim. biophys. Acta (Amst.) **303**, 1 (1973).

Augusteyn, R. C., McDowall, M. A., Webb, E. C., Zerner, B.: Biochim. biophys. Acta (Amst.) **257**, 264 (1972).

Bahl, P. P., Smith, E. L.: J. biol. Chem. **240**, 3585 (1965).

Barnabas, J., Goodnow, M., Moore, G. W.: Comp. Biochem. Physiol. **39**B, 455 (1971).

Bitar, K., Vinogradov, S. N., Nolan, C., Weiss, L. J., Margoliash, E.: Biochem. J. **129**, 561 (1972).

Chan, S. K.: Biochim. biophys. Acta (Amst.) **221**, 497 (1970).
Chan, S. K., Margoliash, E.: J. biol. Chem. **241**, 407 (1966a).
Chan, S. K., Margoliash, E.: J. biol. Chem. **241**, 335 (1966b).
Chan, S. K., Tulloss, I., Margoliash, E.: Biochemistry **5**, 2586 (1966).
Fitch, W. M.: J. molec. Biol. **26**, 499 (1967).
Fitch, W. M.: Systematic Zool. **20**, 406 (1971).
Fitch, W. M.: Haematologie u. Bluttransfusion **10**, 177 (1972a).
Fitch, W. M.: Brookhaven Symp. in Biol. **23**, 186 (1972b).
Fitch, W. M., Margoliash, E.: Biochem. Genet. **1**, 65 (1967).
Fitch, W. M., Markowitz, E.: Biochem. Genet. **4**, 579 (1970).
Goldstone, A., Smith, E. L.: J. biol. Chem. **241**, 4480 (1966).
Goldstone, A., Smith, E. L.: J. biol. Chem. **242**, 4702 (1967).
Gurther, L., Horstmann, H. J.: Europ. J. Biochem. **12**, 48 (1970).
Hartigan, J. A.: Biometrics **29**, 53 (1973).
Heller, J., Smith, E. L.: J. biol. Chem. **241**, 3165 (1966).
King, J. L.: Proc. of the Sixth Berkeley Symp. on Math. Stat. and Prob. Berkeley: Univ. California Press 1972.
Kreil, G.: Hoppe-Seylers Z. physiol. Chem. **334**, 154 (1963).
Langley, C. H., Fitch, W. M.: In: Genetic structure of populations, N. E. Morton, ed. University Press of Hawaii, Honolulu. (1973 in press.)
Lederer, F., Simon, A. M., Verdiere, J.: Biochem. biophys. Res. Commun. **47**, 55 (1972).
Lin, D. K., Niece, R. L., Fitch, W. M.: Nature (Lond.) **241**, 533 (1973).
Margoliash, E., Fitch, W. M.: Miami Winter Symposium, W. J. Whelan, ed., p. 33. Amsterdam: North-Holland Publ. Co. 1970.
Margoliash, E., Smith, E. L., Kreil, G., Tuppy, H.: Nature (Lond.) **192**, 1125 (1961).
Matsubara, H., Smith, E. L.: J. biol. Chem. **238**, 2732 (1963).
McDowell, M. A., Smith, E. L.: J. biol. Chem. **240**, 4635 (1965).
Morgan, W. T., Hensley, C. P., Jr., Riehm, J. P.: J. biol. Chem. **247**, 6555 (1972).
Nakayama, T., Titani, K., Narita, K.: J. Biochem (Tokyo) **70**, 311 (1971).
Narita, K., Titani, K.: J. Biochem. (Tokyo) **63**, 226 (1968).
Narita, K., Titani, K.: J. Biochem. (Tokyo) **65**, 259 (1969).
Needleman, S. B., Margoliash, E.: J. biol. Chem. **241**, 853 (1966).
Nolan, C., Margoliash, E. : J. biol. Chem. **241**, 1049 (1966).
Pettigrew, G. W.: FEBS Letters **22**, 64 (1972).
Pettigrew, G. W.: Nature (Lond.) **241**, 531 (1973).
Ramshaw, J. A. M., Richardson, M., Boulter, D.: Europ. J. Biochem. **23**, 475–483 (1971).
Ramshaw, J. A. M., Thompson, E. W., Boulter, D.: Biochem. J. **119**, 535–539 (1970).
Rothfus, J. A., Smith, E. L.: J. biol. Chem. **240**, 4277 (1965).
Sneath, P. H. A.: J. theor. Biol. **12**, 157 (1966).
Sokolovsky, M., Moldovan, M.: Biochem. J. **124**, 783 (1972).
Stevens, F., Glazer, A. N., Smith, E. L.: J. biol. Chem. **242**, 2764 (1967).
Stewart, J. W., Margoliash, E.: Canad. J. Biochem. **43**, 1187 (1965).
Sugeno, K., Narita, K., Titani, K.: J. Biochem. (Tokyo) **70**, 659 (1971).
Thompson, E. W., Laycock, M. V., Ramshaw, J. A. M., Boulter, D.: Biochem. J. **117**, 183 (1970a).

Thompson, E. W., Notton, B. A., Richardson, M., Boulter, D.: Biochem. J. **124**, 787 (1971c).
Thompson, E. W., Richardson, M., Boulter, D.: Biochem. J. **121**, 439 (1970b).
Thompson, E. W., Richardson, M., Boulter, D.: Biochem. J. **124**, 779 (1971a).
Thompson, E. W., Richardson, M., Boulter, D.: Biochem. J. **124**, 783 (1971b).
Vogel, F.: Hum.-Genet. **8**, 1 (1969).
Vogel, F.: J. molec. Evolution **1**, 334 (1972).

Prof. W. M. Fitch
Department of Physiological Chemistry
University of Wisconsin
Madison, Wisconsin 53706
USA