J. molec. Evolution 1, 334—367 (1972)

Non-Randomness of Base Replacement in Point Mutation

F. Vogel

Institut für Anthropologie und Humangenetik, Universität Heidelberg

Received April 29, 1972

Summary. Using information from human hemoglobin variants, different hemoglobin chains, cytochrome c, insulin molecules, labile parts of human γ-globulin, ($\varkappa$-, λ-, and H-chains) and tobacco mosaic virus coat proteins, some aspects of point mutations were examined.

The main results:

1. All recent hemoglobin variants characterized by one amino acid substitution can be explained by one single base replacement. Of the amino acid substitutions in the other proteins, many more can be accounted for in this way than expected if substitution occured at random.

2. Within the human hemoglobin α-, β-, γ-, and δ-cistrons, a number of codons can be excluded.

3. When origin and direction of base replacements are taken into account, transitions cytosine → thymine (C → T) and thymine → cytosine (T → C) turn out to occur much more often than expected if replacements would occur at random. They are also more frequent than the corresponding transitions guanine → adenine (G → A) and adenine → guanine (A → G). This trend can be observed in all cistrons examined. It cannot be explained by obvious biases in the ascertainment of amino acid substitutions. It points to a relationship between mutation and coding, that cannot be explained on the basis of our present knowledge of the molecular processes involved in replication, mutation, repair, and transscription. Transversions, on the other hand (replacements of a purine by a pyrimidine or vice versa) seem to occur at random.

4. There is no evidence for clustering of point mutations in the same or in neighbouring codons of the abnormal human hemoglobin α- and β-cistrons.

Key words: Genetic Code — Codon Exclusion — Point Mutation — Base Replacement — Non-Randomness of Base Replacement — Amino Acid Substitution.

I. The Problem

One of the most important parameters of evolution is the mutation process. Many mutations are point mutations sensu strictiori, i.e. one base pair only in the DNA double helix is changed. Statistical analysis of some aspects of these point mutations has become possible, since the genetic code has been deciphered and the amino acid sequences of many proteins have become known. Analyses were published among others by Beale and Lehmann (1965), Vogel and Röhrborn (1965, 1966), Epstein (1966), Deran-

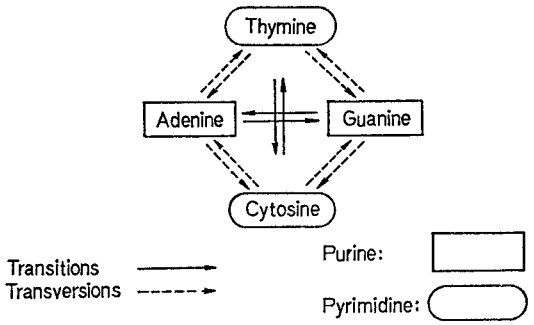


Fig. 1 (quoted from Vogel, 1970). Possible transitions (replacements purine→purine and pyrimidine→pyrimidine) and transversion (replacements purine→pyrimidine and pyrimidine→purine). (From: Chemical mutagenesis in mammals and man; ed. F. Vogel and G. Röhrborn Berlin-Heidelberg-New York: Springer 1970; chapter 2, Fig. 3, p. 31)

court *et al.* (1967), Zuckerkandl *et al.* (1971), Vogel (1969), Fitch (1967, 1972), Lehmann and Carrell (1969), Dellweg (1967), Rahmel (1967).

We had the impression that transitions, i.e. replacements of one pyrimidine by another pyrimidine, or of one purine by another purine, are more than half as common as transversions, i.e. replacements of a pyrimidine by a purine or vice versa (Vogel and Röhrborn, 1965, 1966). Epstein (1966) independently arrived at the same conclusion. He analysed 31 hemoglobin variants and 20 mutations between cistrons for different hemoglobin chains. He not only concluded that transitions are observed more frequently than expected, but that the C→T transition is the most frequent one. He related this result to considerations of Freese and Yoshida (1965) that this result had to be expected on the basis of chemical mutation mechanisms. This conclusion is also contained in Table 11 of the paper of Vogel and Röhrborn (1965). Their Table 10 also shows the high incidence of the C→T transition. In the discussion of this paper, these authors point also to the dilemma that, for example, a C→T transition might indicate a primary event either in C, or in G. Independently, Fitch (1967) has analysed the same problem with hemoglobin mutations and cytochrome *c* replacements. Again, he could show that in the mRNA code, the transition guanine→adenine (G→A) was more frequent than the three others. This corresponded to our result with the human hemoglobin variants (Vogel and Röhrborn, 1965, 1966), that the cytosine→thymine (C→T) transition (expressed in the DNA code) was the most frequent one. We (Vogel, 1969) have reexamined the problem with inclusion of the additional hemoglobin variants known up to that time, and of other hemoglobin mutations. The increased frequency of the C→T transition was confirmed. The most comprehensive analysis of the problem published so far is that of Derancourt *et al.* (1967) and especially Zuckerkandl *et al.* (1971).

These authors not only confirmed the high incidence of the C→T transition. They could also show that the effect is bidirectional in the globins, T→C being also more frequently observed than the corresponding transition A→G. In the other protein examined, the cytochrome c variants, C→T only seemed to be favoured. However, these authors believe that the preferences observed do not reflect any particularity of the mutation process, but selective trends during evolution. They also note a tendency towards substitution of more common amino acids by rarer ones and interpret it as a tendency towards randomness.

Their method of analysis, which deviates from our own in certain aspects, will be discussed later. Suffice it to say that they analysed not only amino acid substitutions, but base replacements, as well. The same was done by Fitch (1971).

Some other problems were also examined by Vogel (1969): Due to the fact that the genetic code is degenerate, no clearcut conclusions are possible from the amino acid sequence to the exact sequence of bases in the cistrons concerned. However, when the assumption is correct that point mutations are due to replacements of one base pair only, some codons may be excluded because they are not compatible with the observed mutations. A number of codon exclusions in the Hb α-, β-, γ-, and δ-chains were tabulated.

Besides, it was shown that the mutations leading to human hemoglobin variants are distributed at random over their cistrons.

In the present paper, it is intended to reexamine some of the problems mentioned, using somewhat more exact methods, and with inclusion of some other proteins. The following problems will be investigated:

1. Which of the point mutations observed are compatible with the assumption that the mutation process has affected one base pair only?

2. Can more codons be excluded than hitherto reported?

3. Do the frequencies of all mutations, transitions and transversions, correspond to expectations which are based on the assumption of randomness, or do they show deviations from randomness? For example, is the C→T transition significally more frequent than expected? Are there any other deviations from the expectation? This problem will be examined with more exact methods than in the past.

4. Moreover, the problem of distribution of mutations over the hemoglobin cistrons of abnormal hemoglobins will be reexamined.

The following data will be included in our study:

1. The 155 human hemoglobin (and myoglobin) variants characterized by one single amino acid substitution, which were listed by H. Lehmann (1972)[1].

1 I am grateful to Dr. H. Lehmann for providing me with this list.

2. The amino acid substitutions noted in comparisons between the hemoglobin β-chains of different species and between β- and δ-chains (Data from Braunitzer, 1967).

3. The differences between sperm whale myoglobin and the human hemoglobin α-, γ-, and β-chains which are thought to have originated from one ancestral chain. Amino acid substitutions were assumed to have occurred in the direction (Mb $\rightarrow\alpha\rightarrow\gamma\rightarrow\beta$). Possible biases introduced by this assumption will be discussed below.

4. Insulin α- and β-chains from different species (see Dayhoff, 1969). Amino acid substitution was assumed in this case to have proceded from the residue more frequently found in the different species examined to the residue found less frequently (or only once).

5. Labile parts of the human gammaglobulin chains ($\varkappa$-, λ-, and H-Chains)[2]. In order to keep as close as possible to the mutation event, only amino acid differences within the subgroups, not between the subgroups were taken into account ($\varkappa$ chains: Subgroups I–III; λ chains: Subgroups I–IV; H-chains: Subgroups I–IV). Again mutation was assumed to have occurred in the direction from the more frequent to the less frequent substitution.

6. Differences between cytochrome c chains of different species (Dayhoff, 1969).

7. Tobacco mosaic virus (TMV) coat proteins (Dayhoff, 1969). Here, TMV vulgare was assumed to be the basic form.

The method of analysis may be demonstrated with one example: In position 57 of the human hemoglobin α-chain we normally find glycine (DNA codons: CCA; CCG; CCT; CCC). In Hb Norfolk, a rare variant, glycine is substituted by aspartic acid (DNA codons: CTA; CTG). This allows us to conclude:

1. The substitution is compatible with only one base replacement in the codon: C $\rightarrow$ T in the second position.

2. One pyrimidine is replaced by another one. Hence, it is a transition.

3. Comparing the possible DNA codons for gly and asp we find out that the replacement C $\rightarrow$ T in the second position will lead to a codon for asp only if the basic codon for gly is CCA or CCG but not if it is CCT or CCC. Therefore Hb Norfolk provides a codon exclusion for CCT and CCC in pos. 57 of the normal human Hb α cistron: The last-mentioned two codons would require two base replacements. For the human Hb variants which are of relatively recent origin this is extremely unlikely; hence the exclusion may be accepted. For the mutations which have been established during evolution, on the other hand, two independent mutational events within the same codon are very well possible. Therefore, we refrained from establishing codon exclusions for these proteins.

---

2 I am grateful to Dr. N. Hilschmann for providing me with alignments available at the end of 1971.

II. Results

1. Point Mutations and Single Amino Acid Replacements

For lack of space, the original tables, which contain all individual amino acid substitutions analyzed, cannot be published. Table 1 shows, how many of the substitutions included in our study can be explained by a single base replacement in the DNA code. It turns out that all human hemoglobin variants analyzed which are due to single amino acid substitutions can be explained by one single base replacement. For the other proteins, however, this is not the case. This can only mean that an appreciable part of these amino acid substitutions are due to at least two independent mutation steps. Theoretically, one could argue that some of them could also be due to molecular events affecting two at the same time. This alternative, however, is rendered very unlikely by the hemoglobin variants: There is no reason why these molecular events should not have affected the hemoglobin cistrons as well. A calculation for the expectations of the percentages of amino acid substitutions which would be compatible with single base replacements, if amino acid substitutions occurred at random, is difficult due to the degeneracy of the genetic code. In order to gain at least a crude impression of the order of magnitude, we calculated how many amino acids can be reached from all codons together of all 20 amino acids. We arrived at the figure of 151 out of 380 ($=20 \times 19$) statistically possible amino acid pairs (39.7%). Furthermore it can be shown that the values for the different amino acids do not differ too much; therefore, the differences in amino acid composition of different proteins may not influence the expectations too strongly. However, this percentage is higher than the

Table 1. Number and percentage of amino acid substitutions which can be explained by single base replacements

	Total number of amino acid substitutions studied	Number of substitutions which can be explained by single base replacements	% of all substitutions
Hb α-variants Hb β-variants (+γ+δ+Mb.)	155	155	100,0
Hb β→s β→δ	28	23	82,142
Mb.→α	107	59	55,514
α→γ	89	55	61,797
γ→β	29	23	79,310
Insulin	43	31	72,093
Human γ	470	374	79,574
Cytochrome *c*	162	97	59,877
TMV coat prot.	127	82	64,567

actual expectations, as it applies for all possible codons of one amino acid together, whereas in every codon only one of the different possibilities can be realized. Therefore, an exact assessment of the expectations is impossible. Suffice it to say that in all proteins examined, amino acid substitutions compatible with the replacement of only one base are much more frequent than expected, if amino acids were replaced at random.

2. Codon Exclusions

The rationale of codon exclusions has been explained above. Tables were published by Vogel and Röhrborn (1965) and Vogel (1969). Due to the additional hemoglobin variants published by Lehmann (1972), the list can be complemented. Table 2 contains an up-to-date list of codon exclusions based on hemoglobin variants. It is remarkable that wherever two independent sources for codon exclusions are available, they confirm each other.

3. Types of Mutations Observed

Table 3 contains the types of mutations observed. Most point mutations can be pinpointed to a certain base replacement. For some mutations, more than one base replacement is possible. For example, asp may have the codons CTA and CTG. glu, on the other hand, may have the codons CTT and CTC. Hence, the substitution asp→glu can be brought about by replacement of one base only. However, it is uncertain which two bases are involved: The following replacements are possible: A→T; A→C; G→T; G→C. All the four are transversions. Therefore, a substitution asp→glu is tabulated as "uncertain transversion". In basically the same way, the uncertain transitions and the completely uncertain base replacements were tabulated.

Statistical analysis of Table 3 renders some interesting results (Table 4).

First, Transitions are again more frequent than expected when replacements would occur at random. For this comparison, calculation of the expected value for transitions posed a certain difficulty: In an earlier paper (Vogel and Röhrborn, 1965), we had assumed that the correct expectation should be 1/3, because every base can undergo one transition and two transversions. Later on, however, we pointed out (Vogel, 1969) that this is not entirely correct, as due to the degeneracy of the code, many replacements, especially transitions, lead to samesense mutations not expressed in an amino acid substitution. The number of visible transitions and transversions can be calculated for every amino acid (Vogel, 1969, Table 9). The correct way to calculate expectations for whole polypeptide chains would be to weigh these expectations with the frequencies of these

Table 2

Nr.	Pos.	+AS	Possible codons	Impossible codons	Informations
			a) Codon exclusion (Hb α-chain)		
1	10	Ala	CGA, CGG	CGT, CGC	Abnormal Hb
2	12	Ala	CGA, CGG	CGT, CGC	Abnormal Hb
3	15	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
4	22	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
5	51	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
6	57	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
7	80	Leu	GAA, GAG, GAT, GAC	AAT, AAC	Abnormal Hb
8	84	Ser	TCA, TCG	AGA, AGG, AGT, AGC	Abnormal Hb
9	92	Arg[a]	GCT, GCC	GCG, GCA, TCT, TCC	Abnormal Hb
10	102	Ser	TCA, TCG	AGA, AGG, AGT, AGC	Abnormal Hb
11	115	Ala	CGA, CGG	CGT, CGC	Abnormal Hb
12	136	Leu	GAT, GAC, GAA, GAG	AAT, AAC	Abnormal Hb
13	141	Arg	GCA, GCG, GCT, GCC	TCT, TCC	Abnormal Hb
			b) Codon exclusion (Hb β-chain)		
1	9	Ser	AGA, AGG	AGT, AGC, TCA, TCG	Abnormal Hb
2	12	Thr	TGA, TGG	TGT, TGC	β-δ
3	14	Leu	GAT, GAC, GAA, GAG	AAT, AAC	Abnormal Hb
4	16	Gly[a]	CCA, CCG	CCT, CCC	Abnormal Hb
5	20	Val	CAT, CAC	CAA, CAG	β-comparison
6	30	Arg	GCG, GCA, TCT, TCC	GCT, GCC	Abnormal Hb
7	32	Leu	GAA, GAG, GAT, GAC	AAT, AAC	Abnormal Hb
8	46	Gly	CCT, CCC	CCA, CCG	Abnormal Hb
9	56	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
10	67	Val	CAA, CAG	CAT, CAC	Abnormal Hb
12	74	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
13	76	Ala	CGT, CGC	CGA, CGG	Abnormal Hb
14	83	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
15	87	Thr	TGT, TGC	TGA, TGG	Abnormal Hb
16	88	Leu[b]	GAA, GAG, GAT, GAC	AAT, AAC	Abnormal Hb
17	91	Leu	GAA, GAG, GAT, GAC	AAT, AAC	Abnormal Hb
18	98	Val	CAC	CAT, CAA, CAG	Abnormal Hb
19	111	Val	CAA, CAG	CAT, CAC	Abnormal Hb
20	113	Val[c]	CAT, CAC	CAA, CAG	Abnormal Hb
21	125	Pro	GGT, GGC	GGA, GGG	β-δ
22	126	Val	CAC, CAT	CAA, CAG	Abnormal Hb
23	129	Ala	CGA, CGG	CGT, CGC	Abnormal Hb
24	136	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
25	141	Leu	GAA, GAG, GAT, GAC	AAT, AAC	Abnormal Hb
			c) Codon exclusion (Hb γ and δ-chain)		
1	γ 12	Thr	TGT, TGC	TGA, TGG	Abnormal Hb
2	δ 16	Ala	CGT, CGC	GGA, GGG	Abnormal Hb
3	δ 22	Ala	CGT, CGC	CGA, CGG	Abnormal Hb
4	δ 136	Gly	CCA, CCG	CCT, CCC	Abnormal Hb
5	δ 116	Arg	GCA, GCG	GCT, GCC	β→δ
			Myoglobin		
1	138	Arg	GCT, GCC	GCA, GCG TCT, TCC	Abnormal Hb

[a] Confirmed by two different substitutions.
[b] Confirmed by comparison between β-chains
[c] Confirmed by amino acid differences between β and δ-chains

Table 3. Base substitution in the transscribed stand

		Hb variants	Hb: $\beta \rightarrow \delta$ subst. + diff. β-chains	Hb: Mb $\rightarrow \alpha \rightarrow \rightarrow \gamma \rightarrow \beta$	Labile γ-Globulin chains	Insulin	Cytochrome *c*	TMV coat proteins	Sum total
T	A→G	9	1	4	17	3	3	3	40
	G→A	8	3	6	15	1	7	5	45
	T→C	22	3	18	71	7	10	11	142
	C→T	32	1	20	58	3	17	11	142
	Uncert.	0	0	0	0	0	0	0	0
	Sum total	71	8	48	161	14	37	30	369
TV	A→T	4	2	4	10	0	5	8	33
	T→A	3	0	5	26	2	3	3	42
	A→C	5	0	11	11	1	2	6	36
	C→A	5	1	7	12	1	3	5	34
	G→T	9	3	6	8	2	4	5	37
	T→G	11	1	9	25	1	9	2	58
	G→C	12	1	12	12	0	4	3	44
	C→G	15	0	8	18	1	6	4	52
	Uncert. transv.	12	5	30	77	8	24	14	170
	sum total	76	14	92	199	16	60	50	507
	Uncertain	4					1	1	

Table 4. Analysis of the data in Table 3

Comparison	Characterization of null hypothesis	Obs. vs. Exp.	χ_1^2	*P*
All transitions vs. all transversions	28,5 % of all base replacements are transitions	Trans.: 369 vs. 249.7	79.72	$\gg 10^{-10}$
Transitions C→T and T→C vs. transitions A→G and G→A	C→T and T→C are as frequent as A→G and G→A	C→T + T→C 284 vs. 184.5	107.32	$\gg 10^{-10}$
Transversions starting with only pyrimidines vs. Transversions starting with purines	same frequency	T + C 186 vs. 168	3.857	0.05
Replacements starting with C or G vs. replacements starting with T or A	same frequency	G + C 354 vs. 352.5	no difference	

amino acid in the polypeptide chains examined. This was done for the hemoglobin α and β chains (Vogel, 1969, Table 10). Applying the same weighting procedure to the polypeptides examined in Table 3 would imply a relatively complicated calculation. Besides, as inspection of Table 9 (Vogel, 1969) shows, differences in expectations between the different amino acids are relatively small. Therefore we decided to carry out these preliminary calculations with the unweighted mean of all expectations for transitions, which is 0.285 (Vogel, 1969, Table 9). The results in Table 4 are based on this expectation.

Secondly, within the group of transitions, those in which the pyrimidines are involved (T→C and C→T) are significantly more frequent than those involving the purines. Earlier studies (Vogel and Röhrborn, 1965; Fitch, 1967; Zuckerkandl, *et al.*, 1971; Vogel, 1969) had pointed primarily towards an increased C→T frequency, and some authors (Fitch, 1967; Vogel, 1969) had pointed out the significance of this result for the understanding of the molecular process of mutation. The present analysis which was carried out with many more proteins and a much greater number of amino acid replacements shows two points: First, the phenomenon is not confined to the hemoglobin and cytochrome c cistrons, but can be shown with a number of other proteins as well. Secondly, not only the C→T, but also the T→C transition is too frequent, i.e. both the transitions in which pyrimidines are involved.

Besides, a small and weakly significant ($P \approx 0.05$) preponderance of pyrimidine involvement can also be observed among the transversions. The effect, however, is much less pronounced than among the transitions. For transitions and transversions together involvement of the T = A base pair (T + A together) is almost exactly as frequent as involvement of the G ≡ C base pair (G + C together).

As will be discussed below, the high frequency of transitions which start with pyrimidines in the transcribed DNA strand confronts us with an intriguing problem about the molecular mechanism of mutations. Therefore, it seems worthwhile to analyse this problem somewhat further.

All mutations for which the base replacement can be ascertained unambiguously affect the first or the second base of the codon. Base replacements in the third position lead to samesense mutations, or to mutations for which the base replacement is ambiguous. For the overwhelming majority of all first and second positions in codons belonging to cistrons for which the amino acid sequence in the polypeptide chain has been analysed, the exact base can be deduced. This is merely another way to restate the well-known fact that the degeneracy of the genetic code is confined almost exclusively to the third bases. Therefore, the number of bases in the transcribed strands of the cistrons analyzed can be calculated almost exactly. This had already been done by Vogel and Röhrborn (1965)

for the codons giving rise to amino acid variants, by Zuckerkandl *et al.* (1971) for hemoglobins and cytochrome *c* and by Fitch (1972) for hemoglobin genes. It will now be done for all the cistrons included in this study. In the exceptional cases in which the first two bases cannot be deduced unambiguosly (codons for leucine, serine and arginine), it will be assumed that all possible codons occur exactly with the same frequency. For leucine, for example, we find the DNA codons AAT, AAC, GAA, GAG, GAT, GAC. This means that the first base is A in two cases and G in four cases. Therefore we count one leu codon as 1/3 A and 2/3 G in the first position. The second position is A in all cases.

The number of bases in the first two codon positions of the cistrons can be compared with the number of mutations affecting these bases, and special hypotheses can be tested directly on the level of the base concerned.

The pooled data are contained in Table 5. Its first four columns contain the percentages of the four bases in the transcribed strands of the cistrons. On the basis of these values, the expectations E in the following columns were calculated. For example, 80 Hb β-variants were analyzed, 42 of which were transitions. The sum of the precentages for T and C together is 54.26%. Thus E for all replacement affecting $T+C = 80.0 \times 0.5426 = 43.41$. E for transitions $(T+C) = 42 \times 0.5426 = 22.79$.

Statistical analysis was carried out using the χ^2 method in the version given by Woolf (1955) with a slight modification:

$$X = \frac{O_1 \times E_2}{E_1 \times O_2}; \quad y = \ln X; \quad V = \frac{1}{O_1} + \frac{1}{O_2}; \quad w = \frac{1}{V}; \quad X_1^2 = y^2 w.$$

Here, O_1 and O_2 are the observed values in the two classes (namely T+C vs. A+G); E_1 and E_2 are the corresponding expectations. The method had been devised by Woolf in order to compare empirical frequencies in two series of patients and controls,—not to compare a frequency with its expectation. In order to take into account this difference, calculation of the variance, which was based originally on four classes (two classes of "patients" and two classes of "controls") had now to be based on the two "observed" classes only. The calculation of χ_1^2 gives—within the limits of rounding—exactly the same results as a classical χ^2 calculation. However, the method used here has two advantages: 1. x gives a reasonable measure for the extent of the deviation from the null hypothesis. 2. Still more important: It is easily possible to calculate a weighted over-all estimate for the deviation from the null hypothesis and χ^2 values for this over-all estimate as well as for the heterogeneity between the different samples examined. For this additional calculation, the following formulas are used:

$$Y = \sum wy / \sum w; \quad Y = \ln X$$

χ_1^2 of deviation: $Y^2 \sum w$; χ^2 of heterogeneity: $\sum wy^2 - Y^2 \sum w$

(Degrees of freedom: Number of single comparisons −1).

Table 5. Expected and observed base replacements in

Protein	% of bases in the transscriding stand (pos. 1 and 2 of codons)				Exspected (E) and observed values for b					
					a) All replacements					
	A	G	T	C	T+C		O/E	X	χ_1^2	P
					E	O				
Hb α-variants	21.65	27.13	26.03	25.06	21.00	28	1.333	2.051	4.587	
Hb β-variants	23.90	21.84	25.75	28.51	43,41	50	1.152	1.405	2.166	
Substitutions Hb $\beta \rightarrow \beta \rightarrow \delta$	23.91	21.84	25.75	28.51	8.68	6	0.691	0.506	1.740	
Myoglobin → Hb α	20.92	21.46	33.12	24.51	27.09	27	0.997	0.993	0	
Hb α → Hb γ	21.65	27.13	26.16	25.06	22.03	25	1.135	1.323	0.825	
Hb γ → Hb β	23.36	19.96	28.18	28.51	13.04	16	1.227	1.747	1.509	
Insulin	32.03	16.99	26.14	24.84	11.22	15	1.337	2.060	2.492	
Human γ-Globulins labile parts of κ, λ and H-chains	26.13	20.77	28.18	24.87	146.00	204	1.397	1.539	45.920	
Cytochrome *c*	17.95	15.71	41.51	28.84	48.43	48	0.991	0.974	0.011	
TMV Coat proteins	23.31	25.95	27.22	23.52	33.49	36	1.075	1.165	0.383	
Sum totals					374.37	455	1.215	1.616	34.413	≈1

χ^2 of heterogeneity: 25.22; P ($m = 9$) = 0.0027.

Table 5 shows the most important results of this analysis:

a) In the total material, replacements affecting T + C are significantly more frequent than replacements affecting G + A. This is also documented by the observation that 8 of 10 values for O/E as well as for X are above 1. However, statistical heterogeneity between the single proteins is also significant, indicating that special conditions influence the extent of the effect.

b) When the analysis is confined to the transitions, the preponderance of T + C compared with A + G is still more pronounced. This turns out also from the higher values of O/E and X, 9 of which are now above 1. In this comparison, the heterogeneity χ^2 is not significant.

c) For transversions, on the other hand, the observed values agree amazingly well with their expectations; heterogeneity is on the borderline of significance.

Taken together, these results show that the risk to be involved in a transition is, indeed, higher for T and C as compared to A and G. A slight and weakly significant increase of transversions involving T and C (see Tables 3 and 4) is not caused by a higher risk of these bases to undergo transversions. It is obviously due to the fact that in the first two codon

'estigated cistrons (1. and 2. bases of the codons only)

lacements starting with T + C as compared with those starting with A + G

Transitions only						c) Transversions only					
+C		O/E	X	χ_1^2	P	T + C		O/E	X	χ_1^2	P
	O					E	O				
1.24	17	1.512	3.246	5.356		9.71	11	1.133	1.313	0.367	
2.79	31	1.360	2.376	6.080		20.62	19	0.921	0.843	0.278	
1.34	4	0.922	0.843	0.058		4.34	2	0.461	0.281	2.417	
3.65	9	1.041	1.103	0.035		18.44	18	0.976	0.945	0.025	
0.24	16	1.562	3.809	5.720		11.78	9	0.764	0.612	1.320	
3.50	13	1.529	4.968	4.454		4.53	3	0.662	0.458	1.144	
7.14	10	1.401	2.404	2.198		4.08	5	1.226	1.603	0.418	
3.49	123	1.439	3.561	41.381		60.53	75	1.239	1.735	7.804	
1.55	27	1.100	1.370	0.723		23.88	21	0.879	0.710	1.025	
3.22	22	1.445	2.670	5.657		18.27	14	0.766	0.618	1.981	
3.16	275	1.388	2.645	60.097	$<10^{-10}$	176.19	177	1.005	1.015	0.018	≈ 0.9

Het) = 11.565; $P(m = 9) \approx 0.5$ χ^2 (Het) = 16.761; $P \approx 0.05$.

positions of the transscribed strands in the cistrons examined, T and C are slightly more frequent than A and G.

An additional question may be examined: Does the increased risk for transitions affect C and T in the same way, or is there any difference between the two bases? This problem was examined for transitions in Table 6. It turns out that C is slightly more often affected than T. (In spite of the fact that the observed numbers are exactly the same. But T is slightly more abundant in the cistrons examined). The difference, however, is not significant.

The mutations leading to hemoglobin variants are of relatively recent origin, whereas others, for example those responsible for the differences between cytochrome c molecules of different spezies are very old. One could expect that peculiarities of the mutation process could become less obvious when the mutations were exposed for a long time to influences of natural selection. The percentage in a protein of amino acid substitutions not compatible with a single base replacement could be taken as one criterion for the ancientness of differences between related proteins. Therefore, we considered it worthwhile to examine the relationship between the ratio observed/expected of the T + C transition (Tables 5 and 6) and the percentage of substitutions not compatible with a single base replacement (Table 1).

Table 6

	T			C		
	E	*O*	*O/E*	*E*	*O*	*O/E*
Hb α-variants	5.727	7	1.222	5.513	10	1.814
Hb β-variants	10.815	12	1.110	11.974	19	1.587
Hb $\beta \rightarrow \beta \rightarrow \delta$	2.060	3	1.456	2.281	1	0.451
Myogl. $\rightarrow \alpha$	4.968	4	1.242	3.677	5	1.360
$\alpha \rightarrow \gamma$	5.232	6	1.147	5.012	10	1.995
$\gamma \rightarrow \beta$	4.227	8	1.893	4.277	5	1.169
Insulin	3.660	7	1.913	3.477	3	0.863
Human γ	45.370	71	1.565	40.041	58	1.449
Cytochrome *c*	15.357	10	0.651	9.191	17	1.850
TMV coat prot.	8.165	11	1.255	7.057	11	1.559
Sum totals	105.581	139	1.317	92.500	139	1.503

Comparison between T and C (from the sum totals). $\chi_1^2 = 2.100$; $P \approx 0.16$.

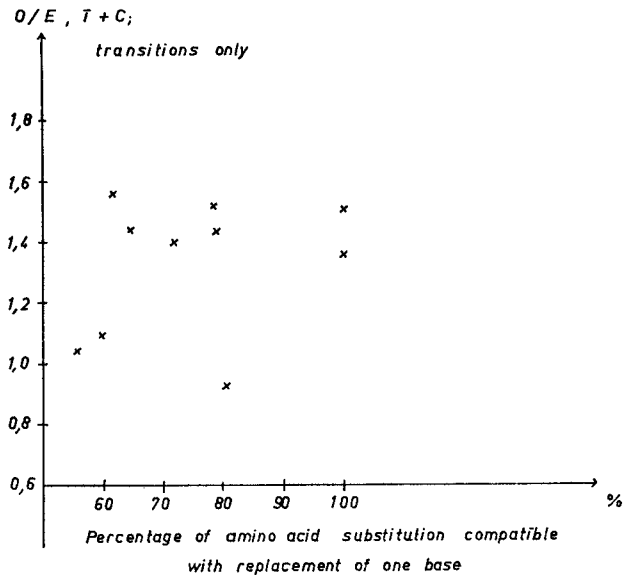


Fig. 2. Relationship between percentage of amino acid substitutions compatible with replacement of one base, and observed/expected number of transitions starting with T or C

The Spearman rank correlation coefficient is, as expected, negative, but the deviation from *O* is negligible, and not significant statistically ($r_s = -0.103$; $p \geqslant 0.05$) (see also Fig. 2). The higher risk of C and T to be involved in transitions as described above could mainly have two different causes: Either, it could be due to biases in the ascertainment of amino acid sub-

stitutions, or it could be due to non-randomness of the mutation process itself.

Which are the ascertainment biases of amino acid substitutions? Two different situations have to be considered: Of the mutations which arose many generations back in evolution, only some of those have been preserved in the genome which were compatible with the function of the protein (Even the majority of these mutations were eliminated by genetic drift, but this does not concern us here). Most of the human hemoglobin variants, on the other hand, have arisen by new mutation only recently. But the probability to be detected is strongly influenced by the character of the substitution involved. Many of them were detected through patients with hematological diseases, and among the others, mutations leading to a change in the electric charge have a much higher change to be detected, as electrophoretic mobility is influenced.

Could the non-randomness demonstrated above simply be caused by non-randomness of the amino acid substitutions ascertained? This seems to be now the opinion of Fitch (1972); and of Zuckerkandl *et al.* (1971).

We compared the observed amino acid substitutions with their expectations. As the possible biases are so different we carried out this analysis separately for the hemoglobin variants on the one hand (confining ourselves to the α- and β-chain variants)—and for the other proteins (leaving out the differences between β- and δ-chains and between the different β-chains). As in the direct analysis of base replacements, we included only mutations involving the first two bases of the codons.

The probabilities P_j for amino acid substitutions were calculated as demonstrated in the following example: For ala, we find C as the first base and G as the second base in the DNA code. C may be replaced by A, G, and T giving the codons for ser, pro and thr respectively. If we assume randomness, these substitutions are expected to occur in the same frequencies, giving expectations of 1/3 each of all replacements to the first base. For the second base, G, the situation is slightly more complicated. It may change into A, T or C. G→A gives the val codon, and G→C gives the gly codon. For G→T, however, there are two possibilities. It depends on the third position of the ala codon, which of the two will occur. If the ala codon is CGA or CGG, G→T will lead to the substitution ala→asp. On the other hand, if the ala codon is CGT of CGC, G→T will lead to the substitution ala→glu. Unfortunately, however, we do not know the third position of the ala codons examined. Therefore, we assume that A and G have the same probability to occur in the third position of the ala codon, as have T and C. This, of course, introduces a small incertainty into our calculation.

We calculate the following values for amino acid substitutions due to base replacements in the second position of this codon:

ala→val 1/3; ala→gly 1/3; ala→asp 1/6; ala→glu 1/6.

Now the results for the first and second base can be combined, giving the p_j values for the different substitutions. They can now be multiplied by the number of alanines observed in the polypeptide chain i under examination (n_i).

In order to obtain expectations $E_{i,j,k}$ for the substitution j of the amino acid k in the polypeptide i, summation over all possible substitutions is

needed. If we call $\sum O_{i,j,k}$ the number of substitutions observed for this polypeptide, the following formula emerges:

$$E_{i,j,k} = \frac{\sum O_{i,j,k}}{\sum\limits_{j,k} n_{ik}\, p_{i,j,k}} \times n_{ik}\, p_{i,j,k}. \tag{1}$$

This calculation may be carried out for all substitutions of each protein i together and additionally for those substitutions which are due to transitions or due to transversions separately.

The ratio $\sum O_{i,j,k}/E_{i,j,k}$ gives an impression of the extent of the deviation from unity. In this way, we calculated the expectations of all amino acid substitutions and for all proteins examined.

The results of this analysis are contained in Table 7. For the hemoglobin variants, the pattern of substitutions observed is clearly nonrandom, the substitutions involving a change in electric charge being much more frequent than expected.

Can the preponderance of transitions in general, or of C→T and T→C transitions be explained by this bias?

As had been shown earlier (Vogel, 1969), there is no difference between all possible transitions and all transversions in the proportion of substitutions with and without change in the electric charge. It can easily be shown that our present calculations carried out in a somewhat different way corroborate this result.

We may, however, ask: Is there any difference between the subgroups A→G; G→A and T→C; C→T?

As inspection of the table shows, there is such a difference, and it points into the direction expected: T→C and C→T replacements lead to a change in the electric charge more frequently than A→G and G→A replacements (see also Zuckerkandl *et al.*, 1971).

65 transitions leading to a substitution with change of the electric charge were observed, 59 of which are T→C or C→T; and 6 are A→G + G→A. The expected ratio

$$\frac{\text{T}\rightarrow\text{C} + \text{C}\rightarrow\text{T}}{\text{A}\rightarrow\text{G} + \text{G}\rightarrow\text{A}}$$

would be 3.2234, whereas the observed ratio is $\frac{59}{6} = 9.8333$. Therefore, this bias does not account fully for the preponderance of T→C + C→T transitions. In the 18 transitions involving no charge difference, the expected ratio

$$\frac{\text{T}\rightarrow\text{C} + \text{C}\rightarrow\text{T}}{\text{A}\rightarrow\text{G} + \text{G}\rightarrow\text{A}} = 0.7191,$$

whereas we observe $8/10 = 0.8$. This result would be compatible with randomness. The numbers, however, are very small.

Calculation of the weighted expectations for the amino acid substitutions not derived from hemoglobin variants was slightly more complicated.

Because all of them were screened by evolution, we decided to calculate expectations for all substitutions $E_{i,j}$ in all proteins and then to combine them to obtain weighted expectations $E_{j,k}=\sum_i E_{i,j,k}$. This means that expectations had first to be calculated for each protein seperately on the basis of the number of substitutions observed in this protein, $\sum O_{ijk}$.

The calculation was carried out in the following way:

$$Ej,k=\sum_i E_{i,j,k}=\sum_i \frac{\sum O_{i,j,k}}{\sum_{j,k} n_{i,k}\, p_{i,j,k}} \times n_{i,k} p_{i,j,k}. \qquad (2)$$

The meaning of the symbols used may be defined once more:

$E_{j,k}$ = Expectation for the substitution j of amino acid k for all proteins together.
$E_{i,j,k}$ = Expectation for the substitution j of amino acid k in the protein i.
$\sum O_{i,j,k}$ = Observed number of all substitutions in the protein i.
$n_{i,k}$ = Number of the amino acid k in the protein i.
$p_{i,j,k}$ = Probalitity for the substitution j of the amino acid k in the protein i.

These expectations were again calculated on the basis of all observed replacements, transitions and transversions together, and additionally for transitions and transversions separately. They are compared with the observed numbers in Table 7.

For these proteins the obvious alternative for the hypothesis of non-randomness in the mutation process itself is the assumption that we mainly observe those amino acid substitutions which involve similar amino acids and, hence, do not influence the properties of the proteins too much. Therefore, it is justified to consider substitutions involving two amino acids in opposite directions together, for example ala→thr and thr→ala. ala→thr means C→T, whereas thr→ala means T→C. This was done in Fig. 3, which shows the ratios O/E for all substitutions with the exception of those that have never occurred. The hatched bars relate to expectations derived from all replacements, transitions and transversions together, whereas the white bars relate to expectations only derived from the total number of transitions or transversions, respectively.

If the substitutions had occurred at random, one would have expected small fluctuations around 1 (dotted line). These are not found, and therefore the whole picture gives a definite appearance of non-randomness.

The next question to be examined is: Which deviations from randomness are most obvious? The answer is easy when we look at the transversions. Here, phe↔tyr and ser↔ala are most frequently involved. These are obviously related amino acids, each pair differing in one OH group only. One can easily imagine that these substitutions may not influence the

Table 7. Expected and observed amino acid substitutions

Amino acid substitution	Polarity +, −	Hb variants					Other proteins (Mb → Hb $\alpha \rightarrow \gamma \rightarrow \beta$; L-chains; Cytochrome C; Insulin; TMV coat prot..)				
		$E_{i,j,k}$		$O_{i,j,k}$	$\frac{O_{i,j,k}}{E_{i,j,k}}$		$E_{j,k}$		$O_{j,k}$	$\frac{O_{j,k}}{E_{j,k}}$	
		T	T+TV		T	T+TV	T	T+TV		T	T+TV
a) Replacements A → G											
Phe → Ser	−	2.30997	1.38523	2	0.86591	1.44378	5.8612	4.0212	2	0.3412	0.4974
Ser → Pro	−	1.58926	0.95305	1	0.62922	1.04926	10.3243	6.7414	4	0.3974	0.5933
Tyr → His	+	2.05372	1.23158	1	0.48692	0.81197	5.6550	3.6913	3	0.5305	0.8127
Cys → Arg	+	0.8507	0.23092	0	0	0	3.1009	2.5596	1	0.3225	0.3907
Leu → Pro	−	6.92991	4.15575	3	0.43291	0.72189	17.2772	11.7926	2	0.1158	0.1696
Leu → Ser	−	3.46496	2.07788	0	0	0	8.6386	5.8962	3	0.3473	0.5088
Met → Thr	−	0.57749	0.34631	0	0	0	2.1470	1.4826	0	0	0
Ileu → Thr	−	0	0	0	0	0	7.4879	5.2152	3	0.4006	0.5752
Val → Ala	−	5.30385	3.18063	1	0.18854	0.31440	11.1143	8.0305	13	1.1700	1.6188
b) G → A											
Ser → Phe	−	0.79463	0.47653	0	0	0	5.1621	3.3707	1	0.1937	0.2967
Ser → Leu	−	0.79463	0.47653	0	0	0	5.1621	3.3707	2	0.3874	0.5933
Pro → Leu	−	1.79700	1.07763	2	1.11130	1.85592	6.2835	4.2149	1	0.1591	0.2373
Pro → Ser	−	1.79700	1.07763	1	0.55648	0.92796	6.2835	4.2149	7	1.1140	1.6608
His → Tyr	+	2.43879	1.46250	5	2.05020	3.41880	2.1876	1.7019	2	0.9142	1.1752
Arg → Cys	+	0.32986	0.19781	0	0	0	1.8877	1.3060	0	0	0
Thr → Met	−	0.68498	0.41077	0	0	0	3.8024	2.5300	1	0.2630	0.3953
Thr → Ileu	−	2.05372	1.23158	0	0	0	11.4297	7.5854	5	0.4375	0.6592
Ala → Val	−	4.62086	2.77105	0	0	0	8.4356	6.0633	1	1.3040	1.8142

c) T → C

Ser → Gly	−	0.79463	0.47653	0	0	0	5.1621	3.3707	10	1.9372	2.9667
Tyr → Cys	−	0.76999	0.46175	1	1.29871	2.16567	5.6554	3.6913	0	0	0
His → Arg	−	2.43879	1.46250	6	2.46024	4.10256	2.1884	1.7019	0	0	0
Gln → Arg	+	0.615992	0.36940	1	1.62339	2.70709	7.6106	5.0468	4	0.5226	0.7926
Met → Val	−	0.57749	0.34631	0	0	0	2.1470	1.4826	6	2.7946	4.0469
Thr → Ala	−	2.73747	1.17986	0	0	0	15.1959	10.1110	20	1.3161	1.9780
Ileu → Val	−	0	0	0	0	0	7.4879	5.2152	16	2.1368	3.0680
Asn → Asp	+	1.28357	0.76974	10	7.79077	12.9913	4.8824	3.3188	5	1.0241	1.5066
Asn → Ser	−	1.28357	0.76974	0	0	0	4.8824	3.3188	17	3.4819	5.1223
Lys → Arg	−	3.38796	2.03170	0	0	0	9.9299	7.0842	21	2.1148	2.9643
Lys → Glu	+	3.38796	2.03170	6	1.77018	2.95319	9.9299	7.0842	6	0.6042	0.8470
Asp → Gly	+	1.92536	1.15461	4	2.07753	3.46437	7.2178	4.8253	3	0.4156	0.6217
Glu → Gly	+	1.84798	1.10820	3	1.62339	2.70709	6.9515	4.9110	7	1.0070	1.4254

d) C → T

Ser → Asn	−	0.74658	0.4471	0	0	0	4.8500	3.1669	11	2.2680	3.4734
Cys → Tyr	−	0.38507	0.23092	0	0	0	3.1010	1.9131	0	0	0
Arg → His	−	0.32986	0.19781	0	0	0	1.8877	1.3060	4	2.1190	3.0628
Arg → Gln	+	0.32986	0.19781	2	6.06317	10.11071	1.8877	1.3060	4	2.1190	3.0628
Arg → Lys	−	0.32986	0.19781	0	0	0	1.8877	1.3060	5	2.6487	3.8285
Val → Ileu	−	3.97908	2.38691	0	0	0	8.9235	6.0247	18	2.0171	2.9877
Val → Met	−	1.32715	0.79587	1	0.75349	1.25648	2.9762	2.0122	4	1.3440	1.9879
Ala → Thr	−	4.62086	2.77105	0	0	0	8.4354	6.0633	21	2.4900	3.4635
Asp → Asn	+	1.92536	1.15461	9	4.67445	7.79484	7.2178	4.8253	15	2.0782	3.1086
Glu → Lys	+	1.84798	1.10820	11	5.95244	9.92600	6.4652	4.6455	6	0.9280	1.2916
Gly → Ser	−	1.81102	1.08604	0	0	0	7.9043	5.2693	9	1.1386	1.7080
Gly → Asp	+	1.81102	1.08604	12	6.62610	11.04931	7.9043	5.2693	4	0.5061	0.7591
Gly → Glu	+	1.81102	1.08604	1	0.55218	0.92078	7.9043	5.2693	8	1.0121	1.5182

Table 7 (continued)

Amino acid substitution	Polarity +, −	Hb variants					Other proteins (Mb → Hb α → γ → β; L-chains; Cytochrome C; Insulin; TMV prot.)				
		$E_{i,j,k}$		$O_{i,j,k}$	$\frac{O_{i,j,k}}{E_{i,j,k}}$		$E_{j,k}$		$O_{j,k}$	$\frac{O_{j,k}}{E_{j,k}}$	
		TV	T + TV		TV	T + TV	TV	T + TV		TV	T + TV
					e) A → T						
Phe → Ileu	−	0.86355	1.38525	0	0	0	3.0314	4.0212	4	1.3195	0.9947
Phe → Tyr	−	0.86355	1.38525	0	0	0	3.0314	4.0212	8	2.6390	1.9895
Ser → Thr	−	0.59412	0.95305	0	0	0	4.7412	6.7414	4	0.8437	0.5933
Tyr → Asn	−	0.29791	0.46184	1	3.47330	2.16525	2.6003	3.6913	4	1.5383	1.0836
Cys → Ser	−	0.14395	0.23092	0	0	0	1.2483	1.9131	0	0	0
Leu → Ileu	−	0.64766	1.03894	0	0	0					
Leu → His	+	1.29533	2.07788	0	0	0	4.4456	5.8963	1	0.2249	0.1696
Leu → Gln	−	1.29533	2.07788	0	0	0	4.4456	5.8963	1	0.2249	0.1696
Met → Lys	+	0.21589	0.34631	0	0	0	0.9364	1.4826	0	0	0
Ileu → Asn	−	0	0	0	0	0	2.6177	3.4776	1	0.3820	0.2876
Val → →Asp	+	0.70353	1.59031	0	0	0	2.9581	4.0152	1	0.3381	0.2491
Val → Glu	+	0.70353	1.59031	3	4.26421	1.88642	2.9581	4.0152	2	0.6761	0.4981
					f) T → A						
Ser → Cys	−	0.29706	0.47653	0	0	0	2.3706	3.3707	0	0	0
Tyr → Phe	−	0.28785	0.46175	1	3.47403	2.16567	2.4123	3.4684	18	7.4618	5.1897
His → Leu	+	0.91171	1.46250	0	0	0	1.5223	1.7019	3	1.9707	1.7627
Gln → Leu	−	0.23028	0.36940	0	0	0	3.6389	5.0468	3	0.8244	0.5944
Arg → Ileu	+	0.06166	0.09891	0	0	0	0.4933	0.6527	0	0	0
Ileu → Phe	−	0	0	0	0	0	2.6717	3.4776	1	0.3743	0.2876
Asn → Tyr	−	0.47985	0.76974	0	0	0	2.4415	3.3188	1	0.4096	0.3013
Asn → Ileu	−	0.47985	0.76974	0	0	0	2.4415	3.3188	4	1.6383	1.2053
Lys → Ileu	+	0.63327	1.01585	0	0	0	2.8124	3.5421	1	0.3556	0.2823
Lys → Met	+	0.63327	1.01585	0	0	0	2.8124	3.5421	1	0.3556	0.2823
Asp → Val	+	0.71977	1.15461	0	0	0	3.5058	4.8253	2	0.5705	0.4145
Glu → Val	+	0.69084	1.10820	2	2.89502	1.80473	3.8832	4.9110	2	0.5150	0.4072

g) A → C

Phe → Val	—	0.86355	1.38525	1	1.15801	0.72333	2.9717	3.5212	2	0.6730	0.5680
Phe → Cys	—	0.86355	1.38525	0	0	0	2.9717	3.5212	0	0	0
Ser → Ala	—	0.59412	0.95305	0	0	0	4.7028	6.7414	18	3.8275	2.6701
Tyr → Asp	+	0.28785	0.46175	0	0	0	2.5754	3.6913	1	0.3883	0.2709
Cys → Gly	—	0.14393	0.23088	0	0	0	1.2483	1.9131	0	0	0
Try → Gly	—	0.43178	0.69263	0	0	0	3.5312	4.9765	0	0	0
Leu → Try	—	0.64766	1.03894	0	0	0	2.174	2.9481	1	0.4614	0.3392
Leu → Arg	+	2.59065	4.15575	4	1.54401	0.962521	2.9700	7.6168	2	0.6734	0.2626
Met → Arg	+	0.21589	0.34631	0	0	0	1.0982	1.4827	1	0.9106	0.6744
Ileu → Ser	—	0	0	0	0	0	3.6078	3.4776	1	0.2772	0.2876
Ileu → Arg	+	0	0	0	0	0	1.3030	1.7376	0	0	0
Val → Gly	—	1.98277	3.18063	0	0	0	5.8278	8.0305	2	0.3432	0.2491

h) C → A

Cys → Phe	—	0.14395	0.23092	0	0	0	1.2483	1.9131	0	0	0
Arg → Leu	+	0.24680	0.39590	1	4.05186	2.52589	1.9472	2.5507	1	0.5136	0.3920
Arg → Met	+	0.06166	0.09891	0	0	0	0.4865	0.6530	0	0	0
Val → Phe	—	0.99138	1.12856	1	1.00869	0.88608	2.9065	4.0020	2	0.6881	0.4998
Ala → Ser	—	1.72745	2.77105	0	0	0	4.6976	6.0632	15	3.1931	2.4739
Asp → Tyr	+	0.71977	1.15461	0	0	0	3.4560	4.8253	2	0.5787	0.4145
Gly → Cys	—	0.67702	1.08604	1	1.47706	0.920776	3.7837	5.3058	0	0	0
Gly → Try	—	0.33851	0.54302	0	0	0	1.8919	2.5343	0	0	0
Gly → Val	—	1.35462	2.17300	1	0.73821	0.46019	7.5707	10.6161	5	0.6604	0.4710
Tyr → Asp	+	0.76775	1.23158	1	1.30250	0.81197	2.5754	3.6914	0	0	0
Ser → Ileu							2.3514	3.3711	3	1.2758	0.8899
Try → Leu							0.5692	0.6702	1	1.7569	1.4921

Table 7 (continued)

Amino acid substitution	Polarity +, −	Hb variants					Other proteins (My → Hb α → γ → L; L-chains; cytochrome C; Insulin; TMV coat prot.)				
		$E_{i,j,k}$		$O_{i,j,k}$	$\frac{O_{i,j,k}}{E_{i,j,k}}$		$E_{j,k}$		$O_{j,k}$	$\frac{O_{j,k}}{E_{j,k}}$	
		TV	T+TV		TV	T+TV	TV	T+TV		TV	T+TV
					i) G→T						
Ser → Tyr	−	0.29706	0.47653	0	0	0	2.1755	3.3707	2	0.9193	0.5933
Pro → His	+	0.33569	0.53949	0	0	0	1.7561	2.1062	0	0	0
Pro → Gln	−	0.33569	0.53849	0	0	0	1.7561	2.1062	1	0.5694	0.4748
Pro → Thr	−	0.67178	1.07763	0	0	0	3.0314	4.2149	1	0.3299	0.2373
His → Asn	+	0.91171	1.46250	0	0	0	1.7285	1.9961	2	1.1571	1.0020
Gln → Lys	+	0.23028	0.36940	1	4.34253	2.70709	3.5892	5.0468	2	0.5572	0.3963
Arg → Ser	+	0.12331	0.19781	0	0	0	0.9730	0.9504	0	0	0
Thr → Asn	−	0.51168	0.82081	0	0	0	3.6015	5.0555	6	1.6660	1.1868
Thr → Lys	+	0.51168	0.82081	0	0	0	3.6015	5.0555	3	0.8330	0.5934
Ala → Asp	+	0.86320	1.38470	4	4.63392	2.88871	2.3474	3.0298	1	0.4260	0.3301
Ala → Glu	+	0.86320	1.38470	0	0	0	2.3474	3.0298	3	1.2780	0.9902
					j) T→G						
Tyr → Ser	−	0.28785	0.46175	0	0	0	2.5754	0.6913	4	1.5532	1.0836
His → Pro	+	0.91171	1.46250	0	0	0	1.7285	1.9961	0	0	0
Gln → Pro	−	0.23028	0.36940	0	0	0	3.5892	5.0468	2	0.5572	0.3963
Thr → Pro	−	1.02336	1.64161	0	0	0	6.8613	10.1109	3	0.4372	0.2967
Asn → His	+	0.47985	0.76974	0	0	0	2.4249	3.3188	2	0.8248	0.6026
Asn → Thr	−	0.47985	0.76974	2	4.16769	2.59827	2.4249	3.3198	7	2.8867	2.1092
Lys → Thr	+	1.26654	2.03170	2	1.57911	0.98440	4.3907	7.0842	5	1.1388	0.7058
Lys → Gln	+	1.26654	2.03170	3	2.36866	1.47660	4.3907	7.0842	7	1.5943	0.9881
Asp → Ala	+	0.71977	1.15461	0	0	0	3.4564	4.8527	5	1.4466	1.0361
Glu → Ala	+	0.69084	1.10820	4	5.79005	3.60946	3.7441	4.9110	7	1.8696	1.4254

					k) G → C						
Ser → Cys	−	0.29706	0.47653	1	3.36632	2.09850	1.7284	3.3707	0	0	0
Ser → Try	−	0.14876	0.23863	0	0	0	1.1775	1.6880	0	0	0
Pro → Arg	+	0.67178	1.07763	3	4.46574	2.78389	3.0314	4.2141	0	0	0
Pro → Ala	−	0.67178	1.07763	0	0	0	3.0314	4.2141	5	1.6494	1.1865
His → Asp	+	0.91171	1.46250	5	5.48420	3.41880	1.4229	1.7019	3	2.1084	1.7627
Gln → Glu	+	0.23028	0.36940	3	13.02671	8.12127	3.5892	5.0468	4	1.1145	0.7926
Thr → Arg	+	0.51168	0.82081	0	0	0	3.6015	5.0555	4	1.1106	0.7912
Ala → Gly	−	1.72745	2.77105	0	0	0	4.6496	5.9795	12	2.5809	2.0069
					l) C → G						
Cys → Ser	−	0.14393	0.23088	0	0	0	1.2483	1.9131	0	0	0
Try → Ser	−	1.43178	0.69263	1	2.31599	1.44377	3.5312	4.9765	0	0	0
Arg → Pro	+	0.24680	0.39590	1	4.05186	2.52589	1.9757	2.6139	0	0	0
Arg → Thr	+	0.12331	0.19781	0	0	0	0.9871	1.3060	4	4.0521	3.0628
Ala → Pro	−	1.72745	2.77105	0	0	0	4.2404	6.0633	7	1.6508	1.1545
Asp → His	+	0.71977	1.15461	6	8.33599	5.19659	3.5058	4.8253	3	0.8557	0.6217
Glu → Gln	+	0.69084	1.10820	3	4.34253	2.70709	3.8832	4.9110	5	1.2876	1.0181
Gly → Ala	−	1.35462	2.17300	0	0	0	7.6983	10.6145	14	1.8186	1.3190

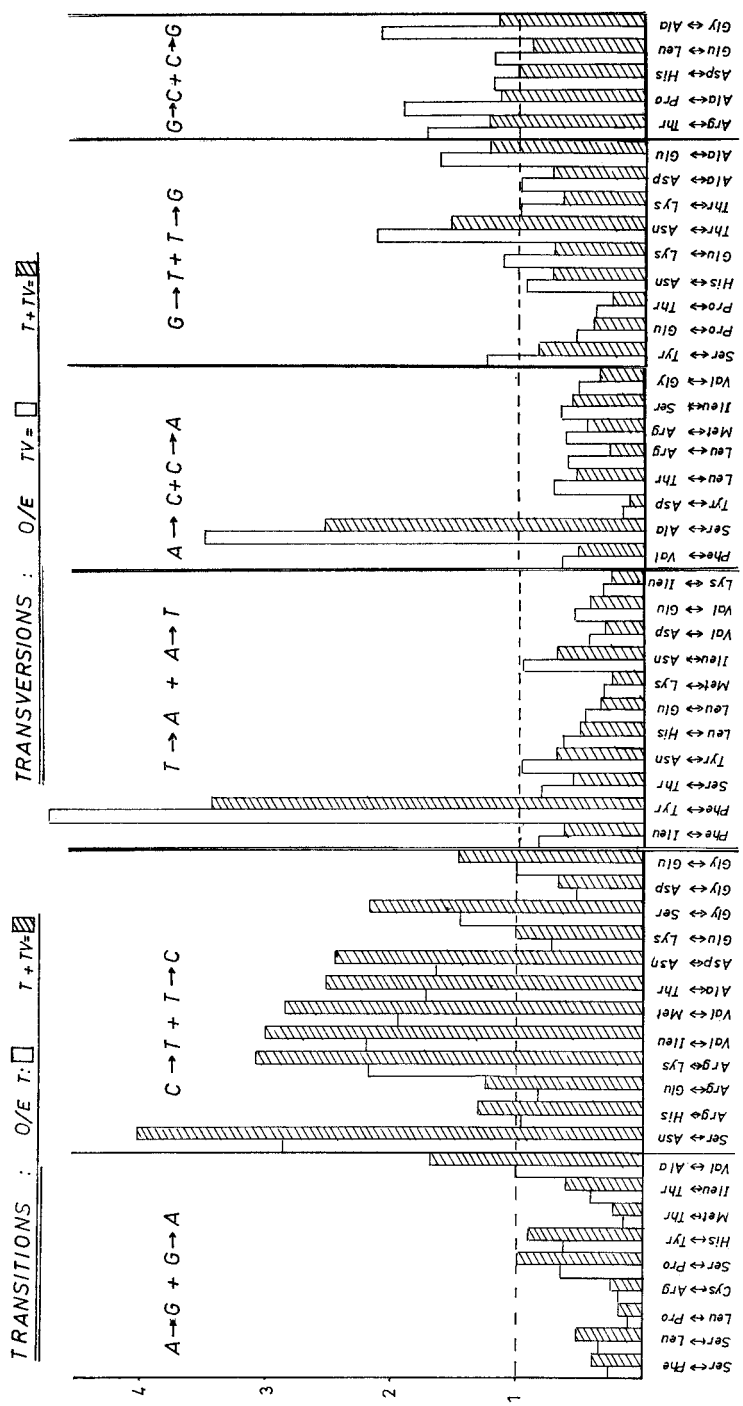


Fig. 3. Comparison of observed and expected frequencies for different amino acid substitutions

properties of a protein too much. One could speculate about the others, especially those that are too rarely involved; however, considering the small sample sizes, the deviations from randomness do not seem to be too obvious.

With the transitions this is quite different. Almost all substitutions involving C→T and T→C transitions are much more frequent than expected. Whereas for some of them, a moderate degree of similarity could be asserted (arg↔lys; val↔ile; asp↔asn), this is not obvious in others (ser↔asn; val↔met; gly↔ser etc.)[3].

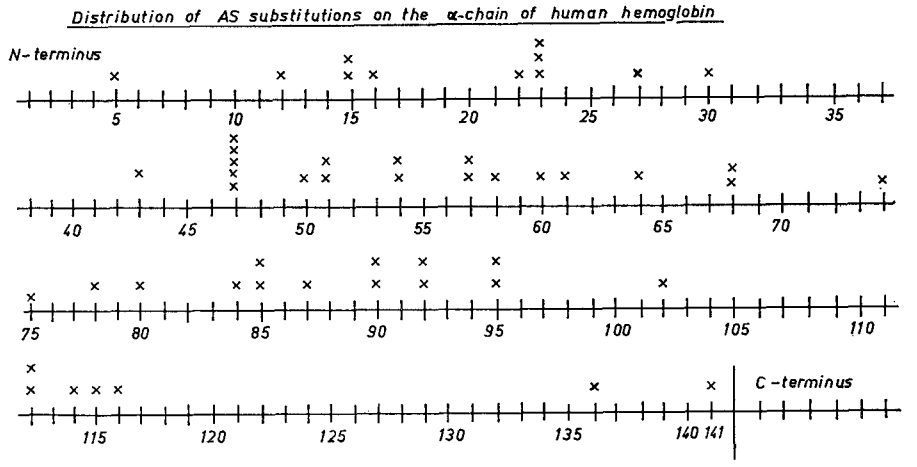
This consistent pattern for the whole group C↔T does not corroborate the assumption that the preponderance of these two transitions is brought about by selection for amino acids which, due to their similarity, can easily substitute each other. On the contrary, it favours the hypothesis that this effect reflects non-randomness of the mutation process itself.

4. The Distribution of Mutations over the Hemoglobin Cistrons

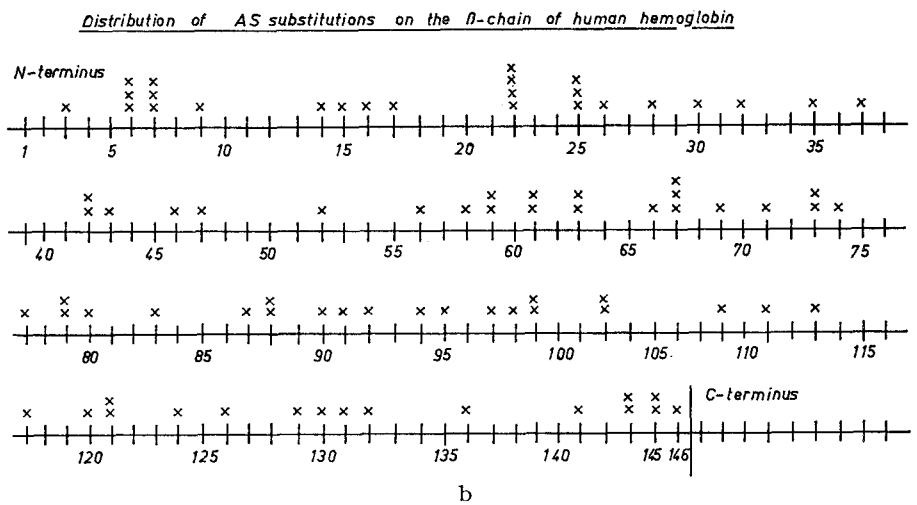
The next problem to be examined is the problem whether the amino acid substitutions known so far are randomly distributed over the length of the α and β chains. The problem has already been examined by Vogel (1969), and the analysis will now be repeated using an increased number of variants. Fig. 4 shows the distribution of 52 α chain variants and 87 β chain variants. First, the problem will be examined whether the number of positions with more than one known substitution is increased as compared with expectation if mutational events occurred at random. The mutations leading to amino acid substitutions are rare events which occur independently of each other. Therefore, the distribution of positions with 0, 1, 2, ... different substitutions is expected to follow a Poisson distribution, if there is a constant probability of mutation. As Table 8 shows, the distributions actually found correspond to the Poisson expectations.

---

3 E. Zuckerkandl (pers. comm.) explained to me: "I think you will meet with general disagreement when you state, implicitly, that some amino acid replacements cannot have been due to amino acid selection, because the amino acids involved in the substitution are too dissimilar. The three examples you mention: Ser↔asn, val↔met, gly↔ser, do not bring out such dissimilarity. If the proportion of molecular sites at which a given substitution has been found to occur is taken as a measure of the degree of functional conservatism of this substitution, then ser↔asn and gly↔ser are highly conservative substitutions, as can be seen for hemoglobins my looking at the matrix published as table 2 in Zuckerkandl and Pauling (1965). The val↔met substitution occurs at a smaller proportion of sites, met being rare, but is also a very "normal" substitution at apolar sites (as can be seen in a table of another paper, given at Berkeley in April 1971, but apparently still not in print). These findings are quite plausible. Asn is like ser a hydrogen bond forming amino acid; ser is like gly a small amino acid; met is like val an apolar amino acid. Concerning the met↔val exchange, it is observed in globins that whenever two different apolar amino acids occur at a given molecular site, the other apolar amino acids of intermediary size (if there are any) will also be found at that site. The size range tolerated at a site is often quite larger."



a



b

Fig. 4a and b. Amino acid substitutions in human haemoglobin variants: distribution over the genes for Hb α-chain (Fig. 3a) and Hb β-chain (Fig. 3b)

The next question to be examined is whether the known substitutions are distributed at random within the total length of the α- and β-chains, or whether they tend to cluster in special parts of the chains. To put the question more exactly, it is asked whether the number of runs of codons with or without known substitutions corresponds to the expectation of a random distribution, or whether it is decreased. The following formulas were used (Siegel, 1956):

$$E_r = \frac{2n_1 n_2}{n_1 + n_2} + 1; \qquad \sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}.$$

Table 8. Comparison with Poisson expectations

$\varkappa$ (number of different mutations in one codon)	α chain		β chain	
	expected	observed	expected	observed
0	97.333	105	80.120	83
1	36.082	24	48.075	47
2	6.679	10	14.422	12
3	0.823	1	2.885	4
4	0.076	0	0.434	1
5	0.007	1	0.064	0

Here, n_1 and n_2 are the numbers of codons with and without known amino acid substitutions, and r is the number of runs. The result: α-chain: $E_r = 54.617$; $O_r = 54$; $\sigma_r = 4.489$; $\chi = E_r - O_r/\sigma_r = 0.137$; $P = 0.92$. β-chain: $E_r = 72.630$; $O_r = 85$; $r = 5.908$; $\chi = E_r - O_r/r = 2.094$; $P = 0.045$.

The difference for the α-chain is not significant. For the β-chain, the difference is weakly significant, but in the opposite direction: The number of runs is somewhat in excess. The hypothesis of a random distribution of mutations leading to amino acid variants within the α and β cistrons is still the more likely one.

III. Discussion

Our investigation had the following results:

1. The majority of the amino acid substitutions in the polypeptides examined can be explained by one base replacement only. This confirms the conclusion that the one prominent mechanism of point mutation is, indeed, the replacement of only one base pair in the DNA double helix. This conclusion is corroborated by the observation that human hemoglobin variants, most of which are rare, and have presumably originated only a few generations ago, can be explained by a single base replacement without any exception. This had already been stated by Beale and Lehmann (1965), and has been confirmed since then repeatedly with an increasing number of variants.

It is not surprising that for the other polypeptide chains analyzed, a strong minority of amino acid substitutions cannot be explained in this way. Here, repeated and independent point mutations within the same codons provide the obvious explanation. In this connection, it may be worthwhile remembering that even for those substitutions which are compatible with a single base replacement, this event is by no means the only possible explanation. For example, if a glycine codon happens to be CCA, then a first step may lead to CTA (asp), and a second step to CTT (glu),

which could also have arisen from the glycine codon CCT by the single replacement C→T. Authors concerned with the rates of amino acid replacements during evolution have given a formula for calculation of these rates from amino acid substitution data (c.f. Zuckerkandl and Pauling, 1965; Kimura and Ohta, 1971). This formula is supposed to correct for multiple substitutions at the same sites, but it does not use the information available from the frequency of "impossible" substitutions. It is not our intention to refine this formula here. Suffice it to say: We have the feeling, that the formula underestimates the number of multiple substitutions somewhat in the examples given by Kimura and Ohta (1971). Holmquist (1972) in his formal treatment points to the same bias.

For the conclusions in the following respects, the possibility of multiple replacements has always to be kept in mind. Therefore, it is cautious always to compare the results obtained with ancient proteins with those worked out on the recent hemoglobin variants. Certain problems, for example the problem of codon exclusion, simply cannot be investigated with ancient proteins: All formal "exclusions" would be invalidated by the possibility of double substitutions.

2. Codon Exclusions. The number of codon exclusions has been increased appreciably compared with the first exclusions published by Vogel and Röhrborn (1965) and Vogel (1969). In some cases, codons for the same amino acids were excluded in different positions. It can be shown occasionally that, as in microorganisms, also in humans different codons for the same amino acid do occur. For example, glycine in pos. 15, 22, 51, and 57 of the α cistron, and in pos. 56, 74, and 83 of the β cistron must be CCA or CCG. In pos. 46 of the β cistron, however, it must be CCT or CCC. Alanine in pos. 10, 12, and 115 of the α cistron, and in pos. 129 of the γ cistron must be CGA or CGG. In pos. 16 of the δ cistron, on the other hand, it must be CGT or CGC. It would be tempting to speculate on the basis of these data about the relative frequencies of different codons in the cistrons concerned; here, however, special conditions involved in the hemoglobin mutations would have to be taken into account. As soon as the exact base sequence of these cistrons will be known from biochemical evidence, it will be interesting to compare the results. Recent advances in the biochemical field seem to be promising: Spiegelman *et al.* (1971) have described a purified RNA-instructed DNA polymerase (reverse transcriptase) from avian myeloblastosis virus, which proved to be an useful tool for synthetizing DNA complements of a rather wide variety of naturally occuring RNAs. It is tempting to speculate about the analytical consequences if this enzyme could be used to synthetize human Hb cistrons.

Theoretically, inspection of the amino acid substitutions in Table 7 should provide information on the problem whether certain codons, though

able to code for amino acids in in-vitro experiments, occur rarely or never in the human cistrons examined: This could have the consequence that some substitutions, though expected with a relatively high frequency, do not occur. For example, if the codons TCT and TCC for arg did not occur, arg→lys, arg→thr, arg→met, and arg→ileu could not be observed. We tried to analyze the material with this problem in mind, but without convincing results.

3. The Frequencies of Different Base Replacements. As mentioned above, it had been shown in earlier papers that certain transitions were more frequent than expected if base replacements occurred at random. This result had been worked out first for recent hemoglobin variants, and later on for differences between hemoglobin chains as well as for cytochrome c variants. The DNA transition C→T seemed to be more frequent than expected. Fitch (1967) and Vogel (1969) had explained that this result, if confirmed, would have far-reaching consequences for our understanding of the spontaneous mutation process (for explanation see below).

Therefore, possibilities for examining this problem somewhat further had to be looked for. Three ways were feasible: a) Examination of a greater number of proteins, b) Investigation of the problem on the level of replacement probabilities of the four bases, and c) examination of the single amino substitutions separately. The second way was tried by Fitch (1972) and by Zuckerkandl *et al.* (1971). Fitch arrived at the conclusion that the probabilities for replacement are more or less the same for all four bases. Zuckerkandl *et al.* (1971), while finding some indications for nonrandomness, concluded that it was probably due to selection, not to the mutation process itself.

We have examined all three aspects:

a) For the recent hemoglobin variants, we took advantage of the kindness of Dr. H. Lehmann, who provided an unpublished paper with a list which was complete up to the second half of 1971. For the labile parts of the human $\varkappa$-, λ-, and H-chains, Dr. Hilschmann gave us an unpublished list, which also contains the data up to autumn, 1971. It can be questioned whether these chains should be included into the analysis. In our opinion, this inclusion is justified because one of the following three hypotheses now under discussion will very probably be correct: Either the substitutions are due to germ cell mutations in very many different genes which had originated during evolution by gene duplication (Hilschmann, 1969), or they are due to somatic mutation (Jerne, in Günther *et al.*, 1972) or both alternatives occur (Ohno, in Günther *et al.*, 1972). We believe that the first alternative will be basically correct; however, even if the second one were exclusively or predominantly correct, this would not impair the usefulness of these mutations for our analysis, provided only that it is warranted to

assume that molecular mechanisms for mutations in germ cells and in somatic cells are basically identical.

For the other proteins, the Dayhoff atlas (1969) was used. Besides the proteins analyzed, it also contains alignments for a number of other proteins. However, they are so fragmentary that we decided not to use them.

It resulted from our analysis that the higher incidence of C→T transitions could also be found in the other proteins, including even the TMV coat proteins.

Besides, it turns out that the finding applies not only to the C→T transition, but to the T→C transition as well. To put it in a slightly different way: All transitions in which pyrimidine bases are involved are more frequent than expected. This last conclusion has to be qualified by one argument: In many of the polypeptide chains compared, the direction of substitution, i.e. the original composition of the codon, is unknown. It can only be inferred from the present state of related chains. Therefore, a T→C transition could be faked by a C→T transition and vice versa. However, in the recent Hb variants, where the direction of mutation is obvious, the T→C transition is also very frequent. Besides, in the cistrons for the other proteins, the direction of mutation can be deduced in many cases with a high degree of probability.

Therefore, until other arguments are available, we assume that both pyrimidines are, indeed, involved.

b) So far, it was only shown that the bases T and C are much more often involved in transitions than the corresponding bases A and G. This could still have two reasons, a trivial or an interesting one. The trivial one would be that the two bases are simply more frequent in the cistrons analyzed, whereas the interesting one would be a higher probability of the single pyrimidine base to be affected.

For the first two positions of every codon, the base can be determined almost without doubt, the degeneray of the code beeing confined mainly to the third position. As explained above, ambiguities have to be taken into account only for the leucine, serine and arginine codons. This was done assuming equal probabilities for all possibilities. Any errors induced by this assumption are bound to be of a much smaller order of magnitude than the deviations actually found. The null hypothesis would imply that all bases are involved in the mutation process in relation to their respective frequencies. As Table 5 shows, this is not the case. Even when all mutations, transversions and transitions, are analyzed together, a significant preponderance of replacements starting with T or C can be detected. This preponderance becomes much more obvious when the analysis is confined to the transitions. Interestingly enough, here, the statistical heterogeneity between the different polypeptides also disappears. Exclusive analysis of the trans-

versions, on the other had, renders no differences at all between the four bases.

Analyzing the problem on the amino acid basis, we had found a weakly significant increase of transversions affecting T and C compared to those affecting A and G. Now this difference turns out to be due to the somewhat higher frequency of C and T as compared to A and G in the cistrons investigated. Here the "trivial" explanation is obviously correct.

It may be mentioned that the over-all base composition of the genes concerned was also examined. It was assumed for the third positions, as well, that the different possibilities are equally likely. Unlike in the first and second positions, however, this introduces an element of considerable insecurity. Therefore, we did not tabulate the results. Suffice it to say that the ratio

$$\frac{A+T}{G+C}$$

generally is not much higher than 1. At the first glance, this seems to contradict over-all results for human and mammalian DNA

$$\left[\frac{A+T}{G+C} = 1.26-1.53;\quad \text{Bresch and Hausmann (1970), Table 6 and 7}\right].$$

This, however, is easily explained by the well-known fact that the highly repetitive DNA included in these analyses contains a high percentage of $A=T$ pairs.

c) The third way of analysis was the comparison of each of the observed amino acid substitutions with its expectation on the assumption of randomness. The substitutions observed show many different deviations from randomness: Among the hemoglobin variants, substitutions which change the electric charge are predomiant. However, it was shown that this bias explains neither the preponderance of transitions, not the extent of increase of $C \rightarrow T$ and $T \rightarrow C$ transitions as compared with $A \rightarrow G$ and $G \rightarrow A$ replacements.

For proteins screened by evolution, non-randomness was also shown. Among the transversions, two substitutions (phe$\leftrightarrow$tyr and ala$\leftrightarrow$ser) can easily be explained by the similarity of the amino acids involved. The $C \leftrightarrow T$ increase, however, is fairly consistent for most of the amino acids involved, and here, similarity is obvious only in very few of them.

Both results—for the hemoglobin variants and for the other proteins—, again favour the hypothesis that the observed phenomenon reflects, indeed, non-randomness of the mutation process itself. They render it very unlikely that it is caused only by a bias in ascertainment of the mutations.

The alternative would be that the amino acid substitutions involving $C \leftrightarrow T$ transitions show a high degree of hidden similarity which enables

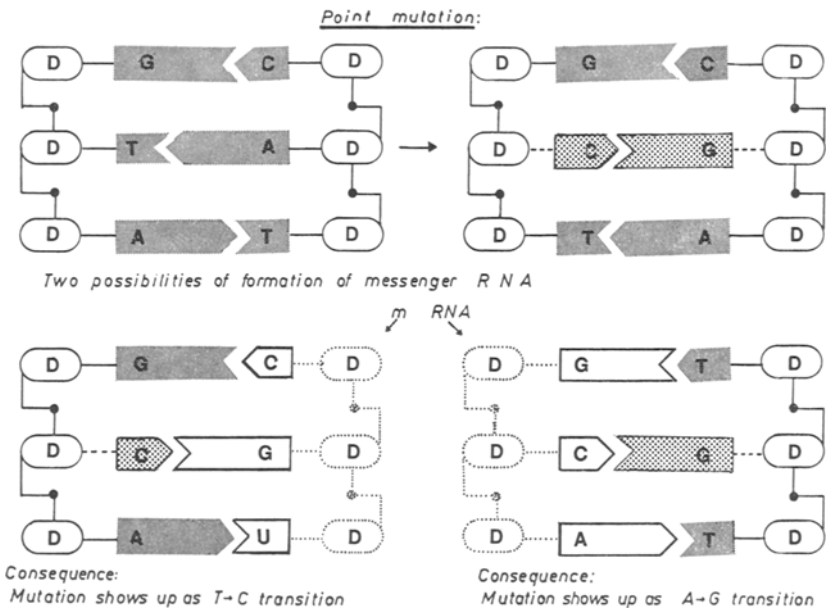


Fig. 5a. A point mutation leads to the replacement of one base pair by an other one. In our example, T = A is replaced by C ≡ G. There are two possibilities for the formation of mRNA: If the left-hand strand is the transscribed one, the mutation will turn up as T→C replacement. If the right-hand strand is the transscribed one, the mutation will appear as A→G replacement. Therefore T→C and A→G replacements would occur in equal frequencies, if the primary process of mutation would be independent form the property (transscribed or complementary) of the strand concerned

them to be substituted in proteins without change of their function. A priori, this hypothesis seems to be somewhat farfetched, but the data in Table 7 and Fig. 3 provide a basis for testing it, taking into account additional properties of the amino acids involved. The remark of E. Zuckerkandl in the footnote of p. 357 represents a beginning of this discussion.

The first two of the three steps of analysis mentioned above were also carried out by Zuckerkandl *et al.* (1971) and by Fitch (1972) with partially different material. These authors, however, proceeded in a slightly different way. We confined the analysis to the amino acid substitutions only involving one base pair replacement, most of which are caused by one step of mutation. The authors mentioned, on the other hand, included also those base replacements which were infered from 'evolutionary trees' constructed in one of the usual ways (for ref. see Zuckerkandl *et al.* 1971). In our opinion, this involves uncertainties which are avoided in our procedure. For practical purposes, this difference in the methods used cannot be too important, as both methods have given similar results.

The main difference concerns the interpretation, and here the third step of our analysis which, to the best of our knowledge, has not been carried

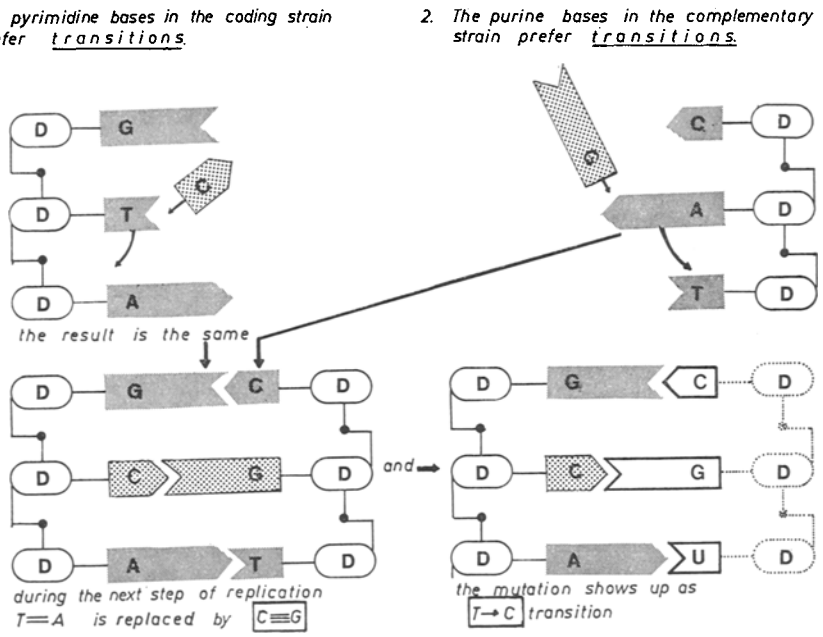


Fig. 5b. The data show that, contrary to expectation, C→T and T→C transitions are much more frequent than the corresponding G→A and A→G transitions. This may have two reasons: 1. The pyrimidine bases (C and T) in the transscribed strand prefer transitions; or 2. The purine bases in the complementary strand prefer transitions. In both cases, the result will be the same: The mutation appears as T→C (or C→T) transition

out before, namely comparison of the expected and observed amino acid substitutions, has given evidence in favour ot the hypothesis that the mutation process itself is involved.

This does not mean, however, that the data do not contain additional evidence for selective processes. On the contrary, this evidence was most clearly established in some cases, the most obvious ones being the preferential preservation of ala↔ser and phe↔tyr substitutions. It would require a special study along the lines given above whether the tendency towards substitution of common amino acids by rarer ones as shown by Zuckerkandl *et al.* (1971) can be confirmed in this way.

Having confirmed the higher probability of C and T bases to undergo transitions with more comprehensive material on the level of the bases concerned, and by separate analysis of each of the amino acid substitutions, we now proceed to the discussion from the point of view of the mutation mechanism.

One would assume without much hesitation that the two properties of the genetic material, mutation and coding, are completely independent from each other. This, however, is not the case (Fig. 5). Replacement of

one base pair (e.g. T = A) by another e.g. C ≡ G) can be initiated either by the replacement of T by C, of A by G. During replication, T will pair with A, whereas C will pair with G. If the primary site of the mutation, which initiates the base pair replacement, were independent from the alternative whether the strand involved is the transscribed or the complementary (non-transscribed) strand, primary mutation events would be quite as frequent in the two types of strands, and a special liability of a base, say C, for a base replacement would lead to an equal increase of C → T and G → A replacements in the transscribed strand, i.e. the cistron under analysis.

This is obviously not so: C is more frequently involved than G, and T is more frequently involved than A.

Whatever we know about the biochemical basis of replication, mutation or repair: There is no hint for an explanation. In this connection, it may be stressed that a hypothesis which would link transcription and mutation directly would not help, as mutation, for example to a hemoglobin variant, occurs in the germ cell, whereas gene activity, in our case hemoglobin synthesis, is confined to a somatic cell many cell generations apart. In order to find an explanation, one would have to look for asymmetries between the transscribed and the complementary strands in the permanent structure of chromosomes or in its replication (different protection by histones?). The question remains open.

4. The last problem to be examined is the question, whether mutations for hemoglobin variants are distributed at random, or whether they show any evidence for clustering in single or neighbouring codons. The extensive data now available favour the assumption that the mutations show, indeed, a random distribution. Hence, we could confirm the result of Vogel (1969) with more comprehensive material. For the other proteins examined, analysis of the same problem did not appear feasible, as these mutations are of very ancient origin, and are strongly selected according to the functional requirements of the proteins concerned.

References

Beale, D., Lehmann, H.: Nature (Lond.) **207**, 259–261 (1965).
Braunitzer, G.: Naturwissenschaften **15/16**, 407–417 (1967).
Bresch, C., Hausmann, R.: Klassische und molekulare Genetik. Berlin-Heidelberg-New York: Springer 1970.
Dayhoff, M. O. (ed.): Atlas of protein sequence and structure, vol. 4. Silver Spring/Maryland: The National Biochemical Research Foundation 1969.
Dellweg, B.: Dtsch. med. Wschr. **92**, 1826–1831 (1967).
Derancourt, J., Lebor, A. S., Zuckerkandl, E.: Bull. Soc. Chim. biol. (Paris) **49**, 577 (1967).
Epstein, C. J.: Nature (Lond.) **210**, 25 (1966).
Fitch, W. M.: J. molec. Biol. **26**, 499–507 (1967).
Fitch, W. M.: In: Haematologie und Bluttransfusionen. München: Lehmann 1972 (in press).

Freese, E., Yoshida, A.: In: Evolving genes and proteins, p. 341. Ed.: V. Bryson and H. J. Vogel. New York: Academic Press 1965.
Günther, E., Albert, E., Kueppers, F., Bender, K: Humangenetik **14**, 173–195 (1972).
Hilschmann, N., Barnikol, H. V., Hess, M., Langer, B., Ponstingl, H., Steinmetz-Kayke, M., Suter, L., Watanable, S.: In: Structure and formation of antibodies. Bayer-Symposium I. Current problems in immunology, p. 69–89. Berlin-Heidelberg-New York: Springer 1969.
Holmquist, R.: J. molec. Evolution **1**, 134–149 (1972).
Kimura, M., Ohta, T.: J. molec. Evolution **1**, 1–17 (1971).
Lehmann, H.: Primärstruktur des Hämoglobins im Verstehen der Funktion. Die Hämoglobinvarianten. (In press.)
Lehmann, H., Carell, R. W.: Brit. med. Bull. **25**, 14–23 (1969).
Rahmel, V.: Mathematische Analyse eines Code-Problems aus der Molekularbiologie. Göttingen: Unpubl. Staatsexamensarbeit 1968.
Siegel, S.: Nonparametric statistics for the behavioural science. New York-Toronto-London: McGraw-Hill Book Company 1956.
Spiegelman, S. Watson, K. F., Kavian, D. L.: Proc. nat. Acad. Sci. (Wash.) **68**, 2843–2845 (1971).
Vogel, F.: Humangenetik **8**, 1–26 (1969).
Vogel, F.: Spontaneous mutations in man. In: Chemical mutagenesis in mammals and man (eds. F. Vogel and G. Röhrborn), p. 16–68. Berlin-Heidelberg-New York: Springer 1970.
Vogel, F., Röhrborn, G.: Humangenetik **1**, 635–650 (1965).
Vogel, F., Röhrborn, G.: Nature (Lond.) **210**, 116–117 (1966).
Woolf, B.: Ann. hum. Genet. **19**, 251–253 (1955).
Zuckerkandl, E., Derancourt, J., Vogel, H.: J. molec. Biol. **59**, 473–490 (1971).
Zuckerkandl, E., Pauling, L.: In: Evolving genes and proteins. V. Bryson and H. J. Vogel, eds., p. 97–166. New York: Academic Press 1965.

Prof. Dr. F. Vogel
Institut für Anthropologie
und Humangenetik der Universität
D-6900 Heidelberg 1
Mönchhofstr. 15A
Federal Republic of Germany