J Mol Evol (1982) 19:20–27

Journal of
Molecular Evolution
© Springer-Verlag 1982

Globular Proteins, GU Wobbling, and the Evolution of the Genetic Code

Jerzy Jurka[1], Zofia Kołosza[1], and Irena Roterman[2]

[1] Department of Tumor Biology, Institute of Oncology, 44–100 Gliwice, Poland
[2] Laboratory of Medical Informatics, Medical Academy, 31034 Kraków, Poland

Summary. It has previously been shown that the formation of GU base pairs in RNA copying processes leads to an accumulation of G and U in both strands of the replicating RNA, which results in a non-random distribution of base triplets. In the present paper, this distribution is calculated, and, using the χ^2-test, a correlation between the distribution of triplets and the amino acid composition of the evolutionarily conservative interior regions of selected globular proteins is established.

It is suggested that GU wobbling in early replication of RNA could have led to the observed amino acid composition of present-day protein interiors. If this hypothesis is correct, then GU wobbling must have been very extensive in the imprecisely replicating RNA, even reaching values close to the critical for stability of its double-helical structure. Implications of the hypothesis both for the evolution of the genetic code and of proteins are discussed.

Key words: GU base pairing – RNA replication – Globular proteins – Genetic code – Evolution

Introduction

As has been discussed by Epstein (1966), Goldberg and Wittes (1966), Volkenstein (1966), and more recently by Wolfenden et al. (1979), the structure of the genetic code evolved to minimize the damaging effects of mutations on protein structures. These statements, however, were based solely on qualitative estimates of the mutation rates which could have influenced evolution of the genetic code towards its present structure (Yčas 1969).

In the preceding paper (Jurka 1977) it was argued that if GU base pairs are allowed during replication of an RNA, then G and U become more abundant in the originating RNA strands than the remaining two bases (i.e. $(G+U)/(A+C) > 1$). Because in the genetic code the non-polar amino acid residues are mainly assigned to codons composed of G and/or U (Volkenstein 1966), it was hypothesized that such a wobbly-replicating RNA could have served as a template for synthesis of polypeptide chains rich in non-polar amino acid residues. The hydrophobic folding, in turn, could have enabled those chains to achieve some compact structures. Furthermore, it was suggested that ancestors of at least some of the modern globular proteins were selected from a set of those proteins. Finally, it was shown that the amino acid composition of the conservative, interior regions of globular proteins qualitatively resembles that expected from the hypothesis.

In this paper we analyse the resemblance statistically and try to assess how extensive the GU wobbling and the associated rate of $A \rightleftharpoons G$ and $U \rightleftharpoons C$ replacements in the replication of RNA could have been.

The previous analysis was performed with the implicit assumption that the genetic code was already in a complete or nearly complete form at the time of the proposed evolution of proteins. We reconsider this assumption in the present paper and discuss the possible influence of GU wobbling for the concomitant evolution of globular proteins and of the genetic code.

Description of the Model

The following scheme represents a wobbly-replicating RNA:

$$C \underset{p_1}{\overset{1}{\rightleftarrows}} G \underset{p_3}{\overset{p_2}{\rightleftarrows}} U \underset{1}{\overset{p_4}{\rightleftarrows}} A$$

Offprint requests to: J. Jurka

0022–2844/82/0019/0020/$01.60

The probabilities of base pairings during the replication are taken as: p_1, p_2, p_3, p_4, or 1. The following two relations hold: $p_1 + p_2 = 1$ and $p_3 + p_4 = 1$.

The GU wobbling during replication of an RNA leads to a generation of transitions, i.e. the following base changes: $A \rightleftarrows G$, $U \rightleftarrows C$.

Let P_{ij} denote the probability of a single transition from a base i to j between plus or minus strands of the RNA. Each of the probabilities defined in this way may be expressed by the above probabilities of base pairings:

$$\begin{aligned}
P_{UU} &= p_4 + p_2p_3 \\
P_{CU} &= p_2 \\
P_{UC} &= p_1p_3 \\
P_{CC} &= p_1 \\
P_{GG} &= p_1 + p_2p_3 \\
P_{AG} &= p_3 \\
P_{GA} &= p_2p_4 \\
P_{AA} &= p_4
\end{aligned} \tag{1}$$

To estimate values of the probabilities of pairing between G and U, the wobble replication of RNA is considered here as a Markov chain. From its properties it follows for the given case that the probabilities of transitions between particular purines or pyrimidines, after n steps (n cycles of replication e.g. strand plus → strand plus), if $n \to \infty$, are convergent. Independently of the initial state a stationary distribution is reached, defined by the following formula:

$$V_j = \sum_i V_i P_{ij} \tag{2}$$

where P_{ij} denotes the probability of transition between bases in a single step (1). From this definition the following stationary distributions for particular bases are obtained:

$$\begin{aligned}
V_A &= \frac{P_{GA}}{1 - P_{AA} + P_{GA}} \\
V_G &= \frac{P_{AG}}{1 - P_{GG} + P_{AG}} \\
V_A &+ V_G = 1 \\
V_U &= \frac{P_{CU}}{1 - P_{UU} + P_{CU}} \\
V_C &= \frac{P_{UC}}{1 - P_{CC} + P_{UC}} \\
V_U &+ V_C = 1
\end{aligned} \tag{3}$$

Furthermore, one may calculate the content of GU pairs (x), formed between complementary strands of the RNA in question, for the stationary distribution:

$$x\,(\%) = \frac{100\, p_2p_3}{p_2 + p_3 - p_2p_3} \tag{4a}$$

and the portion of bases in RNA which undergo transition-type mutations (y):

$$y\,(\%) = \frac{100\, p_2p_3\,(2 - p_2 - p_3)}{p_2 + p_3 - p_2p_3} \tag{4b}$$

Similarly, one can easily calculate the ratios of accumulated (Ac) to intermediate (I) and dissipated (D) triplets for the closed codon groups presented in Table 1.

$$\frac{Ac}{D} = \frac{p_3^2}{p_2^2\, p_4^2} \qquad \text{(Group 1)}$$

$$\frac{Ac}{I} = \frac{I}{D} = \frac{p_3}{p_2p_4}$$

$$\frac{Ac}{D} = \frac{1}{p_1p_4}$$

$$\frac{Ac}{I_k} = \frac{I_\ell}{D} = \frac{p_3}{p_2p_4} \qquad \text{(Groups 2 and 4)} \tag{5}$$

$$\frac{Ac}{I_\ell} = \frac{I_k}{D} = \frac{p_2}{p_1p_3}$$

$$\frac{Ac}{D} = \frac{p_2^2}{p_1^2\, p_3^2} \qquad \text{(Group 3)}$$

$$\frac{Ac}{I} = \frac{I}{D} = \frac{p_2}{p_1p_3}$$

For $p_2 = p_3 = p$ these formulas reach the same simple form for all the four groups:

$$\frac{Ac}{D} = \frac{1}{(1-p)^2} \tag{6}$$

$$\frac{Ac}{I} = \frac{I}{D} = \frac{1}{1-p} \overset{df}{=} \alpha$$

The case in which $p_2 \neq p_3$, although admissable theoretically, is omitted in further analysis because of the lack of supporting evidence for this.

The chi squared minimization procedure gives the following equation for the optimal value of α:

$$2\,D_o^2\,\alpha^3 + (I_{o1}^2 + I_{o2}^2)\,\alpha^2 - (I_{o1}^2 + I_{o2}^2)\,\alpha - 2Ac_o^2 = 0 \tag{7}$$

where D_o, I_{o1}, I_{o2}, Ac_o are the observed numbers of amino acids assigned to dissipated, intermediate and accumulated triplets . Given an optimal value for α and a total number of triplets in a group (equal to the observed total number of amino acids assigned to the group), one may easily calculate the expected numbers of coding triplets.

Analysis of the Amino Acid Composition

General Analysis

The contemporary globular proteins have specified highly ordered tertiary structures. Far-reaching changes of their sequences without loss of function are possible. Only their three-dimensional structure and the active center has to be preserved including, obviously, those residues which participate in a unique way in regulation or catalysis (Anfinsen 1973). Sequences mainly responsible for the maintenance and formation of the tertiary structures are buried in the internal, hydrophobic, densely packed cores. Consequently, residues buried in the interior of globular proteins are carefully preserved by natural selection, unlike those exposed to solvent (Acher 1974). In our further analysis we compare the amino acid composition of the conservative interior regions of globular proteins to the nonrandom distribution of triplets in RNA expected from the model.

As a starting point we consider numbers of the amino acid residues buried over 95% in the interior of nine globular proteins as calculated by Chothia (1975), and discussed qualitatively in the previous paper (Jurka 1977: see Table 1). To distribute leucine between intermediate and accumulated codons, it was assumed that the number of leucines assigned only to the latter ones equals the number of phenylalanines. In this case we have the following summed numbers of the buried residues for all the four groups: $Ac_o = 234$, $I_{o1} + I_{o2} = 271$, $D_o = 78$. The corresponding expected values for $\alpha = 1.73$ are: 234.12, 270.66 and 78.22, respectively. This very good agreement, although reached using a minimal number of assumptions, may be purely accidental. One reason for this possibility is that, we cannot include the nonsense codons in our calculations. To overcome this difficulty we decided to calculate our data in the following way: we remove the octotriplet group containing the nonsense codons and the related ones coding for tryptophane, glutamine and, presumably, one third of the total number of arginines. Also, Also, we consider it more objective to distribute leucine in a manner similar to that for arginine i.e. 1/3 to UUA and UUG, and the remainder to intermediate triplets. This gives a somewhat worse coincidence with the expectations: $Ac_o = 215$ (222.60), $I_{o1} + I_{o2} = 281$ (266.58) and $D_o = 73$ (79.81). The expected numbers are given in brackets ($\chi^2 = 1.62$ for one degree of freedom).

In the same manner we distribute in Table 3 the numbers of residues listed in Table 2, for another set of 15 proteins. The 15 proteins are: ribonuclease S, staphylococcal nuclease, subtilisin novo, subtilisin BPN', papain, glyceraldehyde-3-phosphate dehydrogenase, thermolysin, flavodoxin, elastase, carbonic anhydrase B, carbonic anhydrase C, trypsin, α-chymotrypsin, lactate dehydrogenase, triose phosphate isomerase. We analyse the data for variously defined protein interior, as explained in Table 2.

The correlation between observed and expected numbers is notable for the total sequences (A), and less

Table 1. The genetic code

Ac n	UUU UUC	Phe Phe	(29)	UUA UUG	Leu Leu	(?)	UGU UGC	Cys Cys	(16)	UGA UGG	Term Trp	(9)	Ac n
I p	CUU CUC	Leu Leu	(?)	CUA CUG	Leu Leu	(?)	UAU UAC	Tyr Tyr	(13)	UAA UAG	Term Term		I p k
I n	UCU UCC	Ser Ser	(?)	UCA UCG	Ser Ser	(?)	CGU CGC	Arg Arg		CGA CGG	Arg Arg	(0)	I p l
D n	CCU CCC	Pro Pro		CCA CCG	Pro Pro	(16)	CAU CAC	His His	(8)	CAA CAG	Gln Gln	(5)	D p
			Group 3						Group 4				
Ac n	GUU GUC	Val Val		GUA GUG	Val Val	(91)	GGU GGC	Gly Gly		GGA GGG	Gly Gly	(60)	Ac n
I n k	AUU AUC	Ile Ile		AUA AUG	Ile Met	(69) (14)	AGU AGC	Ser Ser	(?)	AGA AGG	Arg Arg	(0)	I p
I n l	GCU GCC	Ala Ala		GCA GCG	Ala Ala	(71)	GAU GAC	Asp Asp	(17)	GAA GAG	Glu Glu	(13)	I p
D p	ACU ACC	Thr Thr		ACA ACG	Thr Thr	(32)	AAU AAC	Asn Asn	(12)	AAA AAG	Lys Lys	(5)	D p
			Group 2						Group 1				

Eight "closed" octotriplet groups distinguished after Wittman (1962), are put into four bigger units. Codons are defined as assumulated (Ac), intermediate (I) and dissipated (D) depending upon the first two bases. The corresponding residues are denoted as polar (p) and non-polar (n). Numbers of buried residues, given in brackets, are taken from Chothia (1975). Total number of leucines equals 57 and of serines 48. For further explanation see the text

Table 2. Numbers of residues buried in 15 globular proteins

A. Total sequences

	Val	Ala	Ile	Met	Thr	Gly	Asp	Glu	Ser	Arg	Asn	Lys	Phe	Leu	Pro	Cys	Trp	Tyr	His	Gln
10	204	152	120	30	56	129	28	25	93	4	32	9	47	144	38	45	29	33	35	16
20	246	188	139	36	77	170	58	41	120	26	55	20	64	178	54	57	49	63	53	35
30	270	213	156	43	98	200	86	67	155	41	77	41	79	202	68	60	57	82	62	50
40	292	238	171	47	125	233	117	86	189	49	103	79	85	212	84	64	63	110	68	67
50	305	259	178	49	141	252	137	104	217	63	130	120	89	220	102	66	63	127	72	99
100	324	308	186	56	202	347	191	145	331	94	187	237	96	245	145	67	65	144	87	144

B. Non-helical sequences

	Val	Ala	Ile	Met	Thr	Gly	Asp	Glu	Ser	Arg	Asn	Lys	Phe	Leu	Pro	Cys	Trp	Tyr	His	Gln
10	161	89	82	22	35	97	21	14	63	3	24	5	30	91	28	37	25	24	23	12
20	191	106	94	24	55	134	45	24	82	16	43	11	43	114	41	45	37	49	34	24
30	210	124	106	29	70	161	67	41	104	22	60	28	55	130	51	47	42	62	38	32
40	229	137	114	32	93	189	87	51	128	28	78	47	60	139	66	50	46	81	42	39
50	238	153	120	33	106	205	100	60	147	37	99	70	64	145	81	52	46	93	44	58
100	255	187	126	38	158	284	138	85	247	64	147	159	69	158	120	53	48	107	55	91

The 15 proteins are listed in the text. The accessible surface areas in folded proteins have been obtained from C. Chothia and J. Janin (personal communication). Residues are successively defined as buried if 10, 20,..., %, or less of their potential accessible surface areas are available to solvent contact. The potential accessible surface areas are taken from Chothia (1976), and α-helical regions from Levitt and Greer (1977). The final computations of this table were done using modified versions of the program written by J. Ninio (personal communication)

Table 3. Observed and expected numbers of residues buried in 15 proteins

Upper exposure of residues to solvent (%)		Ac		$I_i + I_2$		D			χ^2 for one degree of freedom
		observed	expected	observed	expected	observed	expected	α	
10		473.00	475.78	578.67	573.23	170.00	172.66	1.66	0.13
20		536.33	595.33	781.00	783.33	259.00	257.67	1.52	0.02
30		676.33	675.90	964.00	965.58	346.00	344.34	1.40	0.006
40	(A)	744.67	736.14	1132.00	1150.22	459.00	449.30	1.28	0.60
50		785.33	764.70	1259.67	1296.10	565.00	549.19	1.18	2.02
100		915.67	865.18	1587.00	1679.97	358.00	315.52	1.03	10.30
10		355.33	349.05	277.67	390.00	115.00	108.94	1.79	0.34
20		451.00	434.96	510.67	541.92	184.00	163.79	1.30	3.76
30	(B)	516.33	436.56	634.34	673.04	247.00	228.06	1.48	4.59
40		574.33	543.22	741.34	802.24	326.00	298.24	1.35	9.41
50		607.33	567.40	827.34	905.78	400.00	361.49	1.25	13.71
100		713.67	644.64	1076.00	1213.22	639.00	570.82	1.06	31.05

All the accumulated, intermediates, and dissipated residues from Table 2 are added together with the exception of the octotriplet group containing nonsense and related codons. The expected data derived from formulas 6 and 7. (A) – total sequences; (B) – non-helical sequences

evident for the non-helical sequences of the analysed proteins (see Table 3). The reason for the removal of α-helical regions in the latter case was to expose β-structural elements which are predominant in most proteins (Richardson 1975), and which are believed to form the earliest stable protein structures (Orgel 1972; Brack and Orgel 1975; von Heijne et al. 1978).

The values of α, calculated from Eq. 7, diminish progressively with the extension of interior regions (from top to bottom of the Table 3). This reflects a step-wise decline of the accumulated (non-polar) residues, associated with an opposite tendency among the dissipated ones when moving from the interior to exterior regions.

The best agreement with the expectations can be seen for α ϵ (1.28, 1.66), which corresponds to 12–25% of the GU base pairs in the wobbly-replicating RNA (see Eqs. 6 and 4a for $p_2 = p_3 = p$).

Group Analysis

The analysis presented in Table 3 gives only a crude estimate of the correlation between the regular pattern of distribution of triplets predicted from the imprecise replication of RNA, and the amino acid composition of globular proteins. A more detailed analysis should involve proportions of residues assigned to the separate

Table 4. Observed and expected numbers of residues assigned to Group 2

Upper exposure of residues to solvent (%)		Ac obs.	(Val) exp.	I_1 obs.	(Ala) exp.	I_2 (Ile + Met) obs.	exp.	D (Thr) obs.	exp.	α	χ^2 for two degrees of freedom
10	(A)	204	221.9	152	131.3	150	131.3	56	77.7	1.69	13.4
20		246	264.7	188	161.4	175	161.4	77	98.4	1.64	11.5
30		270	289.6	213	185.6	199	185.6	98	119.0	1.56	10.0
40		292	309.2	238	210.3	218	210.3	125	143.1	1.47	7.2
50		305	320.9	259	226.0	227	226.0	141	159.1	1.42	7.7
100		324	332.1	308	265.7	242	265.7	202	212.5	1.25	9.6
10	(B)	161	169.4	89	87.3	104	87.3	35	45.0	1.94	5.86
20		191	194.9	106	107.7	118	107.7	55	59.5	1.81	1.43
30		210	213.6	124	125.6	135	125.6	70	73.9	1.70	0.99
40		229	226.9	137	143.6	146	143.6	93	90.9	1.58	0.41
50		238	235.3	153	155.8	153	155.8	106	103.2	1.51	0.21
100		255	242.5	187	187.9	164	187.9	158	145.7	1.29	4.74

The observed numbers are taken from Table 2

Table 4a. Observed and expected numbers of residues assigned to Group 1

Upper exposure of residues to solvent (%)		Ac obs.	(Gly) exp.	I_i (Asp + Glu) obs.	exp.	I_2 (Ser + Arg)* obs.	exp.	D (Asn + Lys) obs.	exp.	α	χ^2 for two degrees freedom
10	(A)	129	128.9	53	74.2	97	74.2	41	42.7	1.74	13.13
20		170	174.0	99	118.0	146	118.0	75	80.0	1.47	10.11
30		200	209.9	153	164.3	196	164.3	118	128.6	1.28	8.23
40		223	240.2	203	213.2	238	213.2	182	189.3	1.13	3.86
50		252	256.7	241	255.7	280	255.7	250	254.7	1.00	3.00
100		347	345.6	336	382.0	425	382.0	424	422.3	0.90	10.39
10	(B)	97	95.0	35	51.0	66	51.9	29	28.3	1.83	9.40
20		134	132.9	69	84.3	98	84.3	54	53.5	1.53	5.02
30		161	159.9	108	118.0	126	118.0	88	87.0	1.36	1.41
40		189	185.6	138	150.3	156	150.3	125	121.7	1.24	1.37
50		205	197.9	160	179.1	184	179.1	169	162.0	1.11	2.72
100		284	270.2	223	280.9	331	280.9	306	292.0	0.96	16.53

*Total numbers of serines and arginines from the analysed proteins are assigned to Group 1

octotriplet groups presented in Table 1, but due to the degeneracy of the genetic code, only some larger units, like double octotriplet groups from Table 1, can be analysed. Even in this case, however, the analysis must be restrained de facto to Group 2, because of the lack of detailed information concerning the distribution of leucine, serine and arginine within the remaining groups.

The amino acid residues ascribed to Group 2 make up over 47% of the total number of residues that are over 95% buried in the protein interior. This proportion, although gradually decreasing when less buried residues are included, suggests a meaningful role of the residues contained in this group for the formation and maintenance of the protein interior. Also the properties of their side chains indicate a structural rather than a functional role. Therefore, Group 2 is of particular interest for our analysis.

As can be seen from Table 4, there is only a poor correlation between observed and expected numbers of residues for total sequences, but the correlation for the considered group becomes apparent when only nonhelical regions of the 15 proteins are analysed. The most reliable values of α for the analysed group are within the range 1.51–1.81 which corresponds 20–29% of GU base pairs in the wobbly-replicating RNA (Eq. 4a). From our calculations it follows that a similar correlation can be obtained for Group 1 if about 90% of the total numbers of serines and arginines are assigned to this group. The reliable values of α, assessed from Table 4a, are smaller than for Group 2 and are within the range 1.11–1.58. However, it must be noted that the number of arginines is much below the expectations and their lack is compensated by serines.

NAD-binding Domains

A NAD-binding domain is a common structural element in dehydrogenases whose function if to bind nicotinamide adenine nucleotide (NAD). It is a unit of about

Table 5. The amino acid composition of NAD-binding domains

Ac	Phe Leu	5	Cys Trp	3 1
I	Leu	8	Tyr	1
I	Ser	?	Arg	?
D	Pro	7	His Gln	1 2
	Group 3		Group 4	
Ac	Val	34	Gly	30
I	Ile	19	Ser	?
	Met	2	Arg	?
I	Ala	18	Asp Glu	12 6
D	Thr	8	Asn Lys	6 11
	Group 2		Group 1	

Numbers of amino acid residues from the homologous, non-helical regions of NAD-binding domains of ADHase, LDHase, GAPDase are assigned to a simplified scheme of the genetic code (Table 1). The data are taken from Ohlsson et al. (1974), except for the LDHase sequence taken from Taylor (1977). The total number of serines equals 13 and of arginines 4

150 residues folded into a basic structure of six-stranded parallel β-sheet with four helical segments covering it in pairs on both sides of the sheet. With a few exceptions, alcohol dehydrogenase (ADHase), lactate dehydrogenase (LDHase) and glyceraldehyde-3-phosphate dehydrogenase (GAPDHase) have the same order of secondary structures in the polypeptide chains of their NAD-binding domains. The sequence is as follows: βA-αB-βB-αC-βC-βD-αE-βE-α1F-βF. It is contained within the 1–147 residues in GAPHase, 22–164 in LDHase and 193–318 in ADHase (Ohlsson et al. 1974).

NAD-binding domain is of particular interest for our analysis because being common to various enzymes it must have preceded them in evolution (Rossmann et al. 1974, 1975). Although we do not have exact data for a quantitative definition of the domain interior for all the three proteins listed above, the removal of α-helical sequences generally uncovers the interior regions.

Even though there is no apparent resemblance among equivalent sequences from various domains (Ohlsson et al. 1974), the regularities in the overall composition of their interiors are similar to those observed for other proteins (Table 5). The optimal value of α for Group 2 equals about 2 and resembles critical wobbling in RNA (one GU per three base pairs – see discussion). The α value for Group 1 is smaller but it is difficult to assess exactly. Also, the general distribution of residues in this group is more irregular than expected from the model. We do not wish to discuss the irregularities in more detail because of the limited number of residues available for statistical analysis.

The most striking observation is that the number of residues assigned to Group 3 and 4, compared to the total number of residues, is almost twofold smaller than for the proteins from Table 2.

Discussion

As can be seen from the above analysis, it is possible to find some correlations between the distribution of the base triplets in RNA, expected from the model of wobbly replication, and the actual amino acid composition observed in the interior sequences of globular proteins. A good agreement between expected and observed data can be seen when all the codon groups are added together (excluding the nonsense and related codons). Also it is seen for residues from the non-helical regions of the analysed proteins, assigned to Group 2. The analysis presented for Group 1 is less reliable because we need additional assumptions regarding the numbers of serines and arginines assigned to this group. For the same reason the remaining two groups (Groups 3 and 4) were not considered in this paper.

In general, we observe more alanines and less arginines than is expected from the model. The abundance of alanines is associated with α-helical regions. The deficiency of arginines in proteins is a more general problem (Jukes 1973a). The observed deviations from the expected distribution weaken the proposed model. However, they could also be considered to result from a selection process superimposed on the original primary amino acid sequences translated from the wobbly-replicating RNA. For example, the ionic amino acid residues of arginine, lysine, glutamic acid and aspartic acid can not be removed from an aqueous environment without severe loss of free energy (Tanford 1980). The selection towards formation of stable folded structures could have then favored sequences with a minimal density of these residues. It could have influenced proportions of residues for Group 1 and caused the observed deficiency of arginines in the protein interior.

One could argue that the proportions of residues buried in the protein interior should be determined simply by the properties of their side chains. However, no significant correlation was found between the numbers of residues buried, and their physicochemical properties as taken from Jungck (1978) (data not shown here).

The optimal values of α for the observed composition, calculated from Eq. 7, are within the range 1 to 2. From our model no GU wobbling in RNA replication is expected if α = 1, whereas α = 2 corresponds to one GU per three base pairs. Thus if the triplet distribution is indeed related to the amino acid composition in the proposed way, then GU wobbling must have been very extensive in the imprecisely replicating RNA. From the empirical rules used by Woese and Fox (1975) to reconstruct a secondary structure of 5S RNA, it follows that one GU per four base pairs gives a stable double-helical

RNA. Most probably, one GU per three base pairs is critical for the stability.

From formula 4b it follows that even for some medium values of α (e.g. α = 1.5) over 26% of bases in the RNA undergo transition-type replacements. This raises the problem of how the evolving systems could tolerate such an enormous rate of transitions. One possibility is that the genetic code could first have evolved to a form which was able to neutralize the effects of these replacements on protein structures. In such a reduced code, codons contained in each of the octotriplet groups from Table 1 could code for only one or a small number of similar amino acid residues. This concept is a version of the idea expressed by Woese (1973), who postulated for an early code: "...classes of "related" amino acids assigned *as a whole* to classes of "related" codons *as a whole*". The living system based on our reduced code could also tolerate the GU base pairing in any of the three codon-anticodon positions which could probably occur at that time (see Woese 1973; Jukes 1973b; Barricelli 1977, 1980). Of course, not all of the possible codon groups, listed in Table 1, must have been present in the reduced code. One may choose only two of them expected from an RNY code (Eigen and Schuster 1978; Eigen and Winkler-Oswatitsch 1981), thus focusing the interest on residues from the bottom part of Table 1 (Groups 1 and 2). These residues are predominant in globular proteins, particularly in NAD-binding domains. One may therefore speculate that only codons contained in Groups 1 and 2 were in use during the period of the extensive GU wobbling.

Further evolution of the primitive code must have been associated with a reduction of the GU base pairing both at the replication and translation level. Here we find a reason for the origin of DNA in which T is used instead of U. If we are right, the GT base pairing was less frequent in the primitive copying processes than the GU one.

A further problem is: how could the evolution of the genetic code have been related to the evolution of globular proteins?

To be biologically active, polypeptide chains must form reasonably stable three-dimensional structures (Janin 1979). Even random copolymers of hydrophobic and hydrophilic residues can form compact structures with a highly hindered intramolecular motion (Bychkova et al. 1980). We consider such compact structures, composed of a small number of different amino acid residues, as a possible model for the early globular proteins.

Let us assume that the first distinction in the primitive coding processes was with respect to the polarity of residues assigned to different octotriplet groups. On the basis of our model we tentatively suggest that the amino acid residue(s) assigned originally to triplets from Group 2 (and perhaps Group 3) were non-polar, whereas the residue(s) assigned to Group 1 were polar. Most probably, no residues were assigned to Group 4 at the time of the extensive wobbling. This assignment could originally prevent at least the most frequent transition-type replacements between polar and non-polar residues.

At the second, more advanced stage the rate of transitions during replication decreased when set of the RNA sequences coding for such proteins was transcribed into a more accurately replicating DNA. A common feature of all the sequences transcribed into DNA was a non-random distribution of codons established due to the GU wobbling in RNA replication. It could have influenced the further assignment of residues with certain properties to particular codons which led to the preservation and the further evolution of the three-dimensional structures of proteins. This was achieved by the retention of the non-polar character of the protein interior, by assigning the most abundant triplets to the least polar residues wherever previous assignments (before the appearance of DNA) did not interfere.

In the further evolution of globular proteins a more delicate balance between polar and non-polar side chains in different regions of globular proteins could have been reached due to the random mutations and natural selection. This view is suggested by the observation that weakly polar residues (Ala, Pro, Gly, Thr, Ser) are replaced more often by the non-polar ones (Cys, Val, Met, Ile, Leu, Phe, Tyr, Trp) in the interior, and more often by polar residues (Arg, Lys, His, Gln, Asn, Asp, Glu) on the exterior of globular proteins (Gō and Miyazawa 1980). Also at that stage, an optimal number of the residues forming hydrogen bonds in the protein interior, which are very important for the maintenance of stable tertiary structures (Chothia 1975; Janin et al. 1978), could have been reached.

Confirmation of the hypothesis put forward in this paper cannot be obtained on the grounds of statistical analysis alone and, for this reason, further discussion of this subject must be left open for the time being.

Acknowledgements. J.J. is indebted to Prof. Dr. Manfred Eigen and to Prof. Dr. Władysław Kunicki-Goldfinger for encouragement and valuable discussions. We would like cordially to thank Dr. John Seddon for help with the preparation of the English version of the manuscript and to all of our other colleagues who gave constructive criticism of this work.

This work was partially supported by the Jakub hr. Potocki Foundation and by a grant from the Max-Planck-Gesellschaft.

References

Archer R (1974) Recent discoveries in the evolution of proteins. Angew Chem Intern Edit 13:185–197

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230

Barricelli NA (1977) On the origin and evolution of the genetic code. I. Wobbling and its potential significance. J Theor Biol 67:85–89

Barricelli NA (1980) A note on uracil-guanine pairing in RNA molecules. Evol Theor 5:29–33

Brack A, Orgel LE (1975) β-structures of alternating polypeptides and their possible evolutionary significance. Nature 256:383–387

Bychkova VE, Semisotnov GV, Ptitsyn OB, Gudkova OV, Mitin YuV, Anufrieva EV (1980) The compact structure of statistical copolymers composed of hydrophobic and hydrophilic amino acid residues. J Mol Biol 14:191–278

Chothia C (1975) Structural invariants in protein folding. Nature 254:304–308

Chothia C (1976) The nature of accessible and buried surfaces in proteins. J Mol Biol 105:1–14

Eigen M, Schuster P (1978) A principle of natural self-organization. Part C: The realistic hypercycle. Naturwissenschaften 65:341–369

Eigen M, Winkler-Oswatitsch R (1981) Transfer RNA, an early gene? Naturwissenschaften 68:282–292

Epstein CJ (1966) Role of the amino acid "code" and of selection for conformation in the evolution of proteins. Nature 210:25–28

Gō M, Miyazawa S (1980) Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. Int J Peptide Protein Res 15:211–224

Goldberg AL, Wittes RE (1966) Genetic code: aspects of organization. Science 153:420 –424

von Heijne G, Blomberg D, Baltscheffsky H (1978) Early evolution of cellular electron transport: molecular models for ferredoxin-rubredoxin-flavodoxin region. Orig Life 9:27–37

Janin J, Wodak S, Levitt M, Maigret B (1978) Conformation of amino acid side-chains in proteins. J Mol Biol 125:357–386

Janin J (1979) The protein kingdom: a survey of the three-dimensional structure and evolution of globular proteins. Bull Institut Pasteur 77:337–373

Jukes TH (1973a) Arginine as an evolutionary intruder into protein synthesis. Biochem Biophys Res Commun 53:709–714

Jukes TH (1973b) Possibilites for the evolution of the genetic code from the preceding form. Nature 246:22–26

Jungck JR (1978) The genetic code as a periodic table. J Mol Evol 11:211–224

Jurka JW (1977) On replication of nucleic acids in relation to the evolution of the genetic code and of proteins. J Theor Biol 68:515–520

Levitt M, Greer J (1977) Automatic identification of secondary structure in globular proteins. J Mol Biol 114:181–239

Ohlsson I, Nordström B, Brändén CI (1974) Structural and functional similarities within the coenzyme binding domains of dehydrogenases. J Mol Biol 38:381–393

Orgel LE (1972) A possible step in the origin of the gentic code. Isr J Chem 10:287–292

Richardson JS (1977) β-sheet topology and the relatedness of proteins. Nature 268:495–500

Rossmann MG, Moras D, Olsen KW (1974) Chemical and biological evolution of nucelotide binding protein. Nature 250:194–199

Rossmann MG, Liljas A, Brändén C-I, Banaszak LJ (1975) Evolutionary and structural relationship among dehydrogenases. In: Boyer PD (ed) The enzymes, New York San Francisco London XI:61–102

Tanford C (1980) The hydrophobic effect: formation of micelles and biological membranes. J Wiley and Sons, New York Chichester Brisbane Toronto

Taylor SS (1977) Amino acid sequence of dogfish muscle lactate dehydrogenase. J Biol Chem 252:1799–1806

Volkenstein MV (1966) The genetic coding of protein structure. Biochim Biophys Acta 110:421–424

Wittmann HG (1962) Proteinuntersuchungen an Mutanten des Tabakmosaikvirus als Beitrag zum Problem des genetischen Codes. Z Vererbungslehre 93:491–530

Wolfenden RV, Cullis PM, Southgate CCF (1979) Water, protein folding, and the genetic code. Science 206:575–577

Woese CR (1973) Evolution of the genetic code. Naturwissenschaften 60:447–459

Yčas M (1969) The biological code. Neuberg A, Tatum EL (eds) North-Holland Publ Co. Amsterdam, London

Received June 18, 1982