ON THE POSSIBLE ORIGIN AND EVOLUTION OF THE GENETIC CODE

THOMAS H. JUKES
Space Sciences Laboratory, University of California, Berkeley, Calif. 94720, U.S.A.

Abstract. The genetic code is examined for indications of possible preceding codes that existed during early evolution. Eight of the 20 amino acids are coded by 'quartets' of codons with four-fold degeneracy, and 16 such quartets can exist, so that an earlier code could have provided for 15 or 16 amino acids, rather than 20. If two-fold degeneracy is postulated for the first position of the codon, there could have been 10 amino acids in the code. It is speculated that these may have been phenylalanine, valine, proline, alanine, histidine, glutamine, glutamic acid, aspartic acid, cysteine and glycine. There is a notable deficiency of arginine in proteins, despite the fact that it has six codons. Simultaneously, there is more lysine in proteins than would be expected from its two codons, if the four bases in mRNA are equiprobable and are arranged randomly. It is speculated that arginine is an 'intruder' into the genetic code, and that it may have displaced another amino acid such as ornithine, or may even have displaced lysine from some of its previous codon assignments. As a result, natural selection has favored lysine against the fact that it has only two codons. The introduction of tRNA into protein synthesis may have been a cataclysmic and comparatively sudden event, since duplication of tRNA takes place readily, and point mutations could rapidly differentiate members of the family of duplicates from each. Two tRNAs for different amino acids may have a common ancestor that existed more recently than the separation of the prokaryotes and eukaryotes. This is shown by homology of two *E. coli* tRNAs for glycine and valine, and two yeast tRNAs for arginine and lysine.

1. Introduction

The evolution of the genetic code has been a favorite subject of speculation for some years. Two general approaches are often used. The first consists of attempts to find an affinity between amino acids and nucleic acid bases. The idea is to show how a primitive translation system may have started prior to the origin of life. This approach has the advantage of giving rise to laboratory experiments. The second approach, which I shall discuss first, is to work backward from the present code to something simpler. One attraction of this procedure is that it is linked to actual existence, which makes an author feel more secure, like a rock-climber who uses a rope on a cliff, rather than 'climbs free'.

The present code (Table I) has much to commend it. One of its good features is that no one was able to guess what it was, although various people tried to do so. This may mean that no-one will be able to guess how the code started.

One obvious feature of the code that makes one speculate on its possible evolution is its pattern of degeneracy. Eight amino acids: alanine, arginine, proline, glycine, leucine, serine, threonine, valine are coded by two bases plus 'any base in the third position'. An example is the code GCN, for alanine, where N is any base. It is tempting to suggest (Jukes, 1966) that an earlier code, from which the present code is descended, may have been for only 15 amino acids, each coded by a doublet and any third base to act as a 'spacer'.

Origins of Life **5** (1974) 331–350. *All Rights Reserved*

TABLE I

The genetic code

UUU Phenylalanine	UCU Serine	UAU Tyrosine	UGU Cysteine
UUC Phenylalanine	UCC Serine	UAC Tyrosine	UGC Cysteine
UUA Leucine	UCA Serine	UAA Chain Termn..	UGA Chain Termn.
UUG Leucine	UCG Serine	UAG Chain Termn.	UGG Tryptophan
CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine
CUC Leucine	CCC Proline	CAC Histidine	CGC Arginine
CUA Leucine	CCA Proline	CAA Glutamine	CGA Arginine
CUG Leucine	CCG Proline	CAG Glutamine	CGG Arginine
AUU Isoleucine	ACU Threonine	AAU Asparagine	AGU Serine
AUC Isoleucine	ACC Threonine	AAC Asparagine	AGC Serine
AUA Isoleucine	ACA Threonine	AAA Lysine	AGA Arginine
AUG Methionine	ACG Threonine	AAG Lysine	AGG Arginine
GUU Valine	GCU Alanine	GAU Aspartic acid	GGU Glycine
GUC Valine	GCC Alanine	GAC Aspartic acid	GGC Glycine
GUA Valine	GCA Alanine	GAA Glutamic acid	GGA Glycine
GUG Valine	GCG Alanine	GAG Glutamic acid	GGG Glycine

An appealing feature of this suggestion is the finding that the same transfer RNA molecule can be the ancestor of two tRNAs, each for a different amino acid. This is shown by comparing the sequences of *E. coli* glycine and valine tRNAs, and of yeast lysine and arginine tRNAs (Table II). Therefore, there is an existing process for switching the translation of a codon from one amino acid to another. However, there is no way of identifying the amino acid that four codons may have specified in a 'earlier code'. For example, were all four codons that start with AG originally assigned to arginine, or to serine, or perhaps to neither one, since arginine and serine each have four other codons?

2. Ambiguous Pairing Between the First Base of Anticodons and the Third Base of Codons

Some anticodons can pair with more than one codon. For example, it was found that the same phenylalanine tRNA would pair with the trinucleotides UUU and UUC; furthermore, no phenylalanine tRNA has been found that will pair with only UUU, or only UUC. When phenylalanine tRNAs were sequenced, the anticodon was found to be GmAA in yeast and wheat (Gm = oxymethyl guanosine) and GAA in *E. coli*. The G or Gm paired with both C and U. This and other findings led Crick (1966) to propose that, while the pairing between the first two bases of the codon and the last two bases of the anticodon followed strictly the regular A·U and G·C rule, there was a certain amount of 'play' or 'wobble' in the pairing between the first base of the anticodon and the third base of the codon. This would be useful biologically, because the utmost fidelity is needed in translating the first two bases of the codon, which are

TABLE II

Comparison of base sequences (unmodified) of *E. coli* tRNAs Valine$_{2B}$ and Glycine$_3$, and Yeast tRNAs Lysine$_H$, and Arginine$_{3A}$

Val EC2B	G C G U U C A	U A	G C U C	A G U U - G G U - U A	G A G C	A	C C A C C	U U G A C A U
Gly EC3	G C G G G A A	U A	G C U C	A G U U - G G - - U A	G A G C	A	C G A C C	U U G C C A A
Val EC2B	G G U G G	G G G - C	G G U G G	U U C G A G U	C C A A U	U G A A C G C	A	% Identical bases = 74
Gly EC3	G G U C G	G G G U C	G C G A G	U U C G A G U	C U C G U	U U C C C G C	U	
Lys YH	G C C U U G U	U G	G C G C	A A U C - G G - - U A	G C G C	G	U A U G A	C U C U U A A
Arg Y3A	G C G C U C G	U G	G C G U	A A U - - G G - - C A	A C G C	G	U C U G A	C U U C U A A
Lys YH	U C A	U A A G N U U	A G G G G	U U C G A G C	C C C C U	A C A G G G C	U	% Identical bases = 73
Arg Y3A	U C A	G A A G A U U	A U G G G	U U C G A C C	C C C A U	C G A G U G C	G	

The vertical lines separate helical and non-helical regions.

strictly specific for 8 of the amino acids (see Table I), but the third base of the codon is ambiguous or 'degenerate' in the codes for 18 of the amino acids. Therefore, it might be 'biologically economical' to use the same tRNA for both phenylalanine codons.

By building models, Crick concluded that (i) the alanine anticodon IGC, in which adenine (in adenosine) has been deaminated to hypoxanthine (in inosine), would pair with GCC, GCU and GCA. (ii) either of the two chainterminating codons UAA and UAG would pair with the anticodon UUA. This is thought to be in an *E. coli* serine suppressor tRNA as a result of a mutation from UGA to UUA in the anticodon. (iii) the anticodon CUA, found in another serine suppressor tRNA as a result of a mutation from CGA, would pair only with UAG. Therefore, I would 'wobble' with C, U or A; U would 'wobble' with A or G, but C in the first anticodon position would be specific for G in the third position in the codon.

True to prediction, inosine has been found in the anticodons only of tRNAs for amino acids that have three codons containing U, C and A in the third position. They were alanine, serine, isoleucine, valine and arginine. Most of these tRNAs are from eukaryotic organisms, but one for arginine is from *E. coli,* the tRNAs of which, in bulk, contain very little I.

A new development was reported by Murao *et al.* (1970). They found that uridine-5-oxyacetic acid in the first position of the anticodon oacUAC in *E. coli* valine tRNA would pair with A, G and U, present in valine codons GUA, GUG and GUU, so that the pairing of G in codons was not specific for C in the anticodon. Furthermore, and in contrast, a number of anticodons were found to contain 2-thiouridine in position 1, pairing only with A (Nishimura, 1972), rather than with A and G as in the case of unmodified U (Table III). Evidently, the first base in the anticodon can be modified to produce either greater or less pairing specificity. I conclude that codon-anticodon pairing has changed during evolution to the extent that modification of the bases in anticodons has altered the specificity of pairing with codons in the directions both of more and less ambiguity.

3. The Importance of U·G Pairing in the Evolution of the Code

A striking thing about the code is that eleven amino acids are coded by pairs of codons that end either in a pyrimidine or in a purine, but no amino acid is coded only by a pair of codons one of which ends in a pyrimidine and the other in a purine. To give an example, histidine has the codons CAU and CAC; glutamine, CAA and CAG. Surely this pattern must have some evolutionary significance, because the pattern is a functional one, and is related to codon-anticodon pairing. Codons ending with a pyrimidine will pair with anticodons starting with G. Codons ending with a purine will pair with anticodons starting with U. These two general rules hold except for anticodons that start with modified bases. Such modifications are probably of recent evolutionary origin since they differ in different species.

Crick's proposal shows that this pattern in the code depends on U·G pairing between the first base in the anticodon and the third base in the codon ('1–3 pairing')

TABLE III

First base of anticodons in sequenced tRNAs compared with third base of codons paired

First base in anticodon	Pairing with	Organisms (E = eukaryotes P = prokaryotes)	In tRNAs for
H (I)	U, C, A	E, P	Ile, Val, Ser, (UCN), Ala, Arg (CGN)
oa^{5}U	U, A, G	P	Val, Ser (UCN)
G	C, U	E, P	Phe, Leu (CUN), Ile, Val, Tyr, Asp, Arg (CGN), Gly
Gm	C, U	E	Phe
G† (‘Q’)	C, U	P	Tyr, His, Asn, Asp
U	A, G	P	Gly, Term.
s^{2}U	A	E, P	Gln, Lys, Glu
U†	A, G(?)	E, P	Leu (UUR), Arg (AGR)
C	G	E, P	Leu (CUN, UUR), fMet, Gln, Lys, Trp, Gly
ac^{4}C	G	P	Met
Cm	G	E	Trp

Abbreviations: H, hypoxanthine; I, inosine; oa^{5}U, uridine-5-oxyacetic acid; Gm, 2′-0-methylguanosine; G†, ‘Q-base’ (an unidentified guanosine derivative); s^{2}U, 2-thiouridine; U†, unidentified uridine derivative; ac^{4}C, 4-acetylcytidine; Cm, 2′-0′methylcytidine; N, unidentified nucleoside; R, purine.

as follows:

→		
Anticodons	GUG	UUG
Codons	CAC	AAC
←	UAC	GAC
	His	Gln

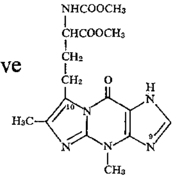
We shall now examine the possibility for U·G pairing between the *third* base in the anticodon and the *first* base in the codon in an earlier genetic code (‘3–1 pairing’). It appears that an elaborate system to *prevent* this ambiguity has evolved in tRNA. This is expressed in special modifications of the nucleoside next following the anticodon (site 40) in tRNA (Table IV). There is, in most cases, an isopentenyl-containing side-chain at site 40 when there is 3–1 pairing between A and U, and a threonyl-containing side-chain at this location when there is 3–1 pairing between U and A, and, in prokaryotes, a methyl group when there is 3–1 pairing between G and C. An obvious role for such side-chains is to prevent 3–1 pairing between G and U, and U and G respectively. The prevention of this pairing is necessary to ensure fidelity of translation of codons starting with U or A, *except* in the case of the initiation of polypeptide chains. In *E. coli* the initiator codon is AUG or GUG, either of which is translated as methionine by

TABLE IV

Modifications of the nucleosides at site 40 in tRNAs as correlated with recognition of codons in *E. coli* and other prokaryotes (Prok.) and yeast and other eukaryotes (Euk.)

Nucleoside adjoining Anticodon			Nucleoside adjoining Anticodon			Nucleoside adjoining Anticodon			Nucleoside adjoining Anticodon		
Codon	Prok.	Euk.	Codon	Prok.	Euk.	Codon	Prok.	Euk.	Codon	Prok.	Euk.
UUY	ms^{2}i^{6}A	*Y*-base	UCY		i^{6}A	UAY	msi^{6}A	i^{6}A	UGY	ms^{2}i^{6}A	
UUA	ms^{2}i^{6}A		UCA	ms^{2}i^{6}A	i^{6}A	UAA			UGA	–	
UUG	ms^{2}i^{6}A	m^{1}G	UCG	ms^{2}i^{6}A		UAG			UGG	ms^{2}i^{6}A	A
CUY	m^{1}G		CCU			CAY	m^{2}A		CGY	m^{2}A, G	A
			CCC								
CUA			CCA			CAA	m^{2}A		CGA	m^{2}A	A
CUG	m^{1}G		CCG			CAG	m^{2}A		CGG		
AUY	t^{6}A	t^{6}A	ACU			AAY	t^{6}A		AGY	t^{6}A	
			ACC								
AUA		t^{6}A	ACA	mt^{6}A, t^{6}A		AAA	t^{6}A	t^{6}A	AGA		t^{6}A
AUG	t^{6}A		ACG			AAG	t^{6}A	t^{6}A	AGG		
AUG-f	A	t^{6}A									
GUY	A	A†	GCY		meI	GAY	m^{2}A	m^{1}G	GGY	A	
GUA	m^{6}A	A†	GCA	A	meI	GAA	m^{2}A		GGA		
GUG	m^{6}A		GCG			GAG	m^{2}A		GGG	A	

Y-base, a guanosine derivative:
m^{1}G, 1-methylguanosine
A†, unidentified adenosine derivative
mel, 1-methylguanosine
m^{2}A, 2-methyladenosine
m^{6}A, 6-methyladenosine
mt^{6}A, N-methylcarbamoyl in t^{6}A



i^{6}A = 6-(Δ^2-isopentenyl) adenosine
t^{6}A = N-(9-(β-D-ribofuranosyl) purin-6-ylcarbamoyl) -L-threonine[a]
[a] Threonine-containing nucleosides were found in tRNAs accepting arginine and serine
Y = U or C
ms^{2}i^{6}A, 2-methylthio-6-(Δ^2-isopentenyl) adenosine

a special tRNA with the anticodon CAU but *without* modification of the adenosine at position 40.* All such modifications take place by enzymatic mechanisms following transcription. These mechanisms can be presumed to be evolutionary innovations. Prior to their appearance, U·G and G·U 3–1 pairing could have taken place in the primitive prokaryotes. Under these conditions, codons starting with G would pair with anticodons ending with either C or U. Therefore the valine codons, GUN, would pair with the eight anticodons represented by NAC and NAU. Similarly, the leucine anticodons represented by NAG would pair with eight codons consisting of CUN and UUN. According to this relationship, isoleucine and phenylalanine would not have been represented in the code. By applying this model throughout codon-anticodon pairing, a code containing ten amino acids is found to be possible (Table V). Eight more amino acids would become members of the code as a result of the suppression of U·G and G·U 1–3 pairing (Table Vb). Two chain terminator codons are introduced

* Note, however, that in eukaryotes this adenosine is modified, and the initiator tRNA pairs with both AUG and GUG.

TABLE V

(a) The ancestral code for ten amino acids, showing anticodon-codon pairing in the first position of codons that includes ambiguity produced by pairing between G and U in addition to G·C and A·U pairing; (b) code 2, produced as a result of elimination of this ambiguity (c) present code resulting from changes in (b). N = U, C, A or G; R = A or G; Y = U or C.

(a) Ancestral Code:

→		NAG					RUG	YUG				
Anticodons	NAU	NAA	NAG	NAC	RUU	YUU	RUA	YUA	RUG	YUG	RUC	YUC
Codons	NUA	NUU	NUC	NUG	YAA	RAA	YAU	RAU	YAC	RAC	YAG	RAG
←	NUG				YAG	RAG						
Amino acid	Val	Leu	Leu	Val	Asp	Glu	His	Gln	His	Gln	Asp	Glu

→		NGG				NCG		
Anticodons	NGU	NGA	NGG	NGC	NCU	NCA	NCG	NCC
Codons	NCA	NCU	NCC	NCG	NGA	NGU	NGC	NGG
←	NCG				NGG			
Amino acid	Ala	Pro	Pro	Ala	Gly	Arg (Orn)	Arg (Orn)	Gly

(b) Changes in (a) leading to code 2 as a result of enzymatic modification of tRNA base 40 and consequent suppression of U, G pairing.

→ Anticodons	NAU	RAA	YAA	NAG	NAC	RUU	YUU	RUA	YUA	RUG	YUG	RUC	YUC
Codons ←	NUA	YUU	RUU	NUC	NUG	NAA	RAA	YAU	RAU	YAC	RAC	YAG	RAG
Amino Acid	Ile	Phe	Leu	Leu	Val	Asn	Lys	Tyr	C.T.	His	Gln	Asp	Glu

→ Anticodons	NGU	NGA	NGG	NGC	NCU	NCA	NCG	NCC
Codons →	NCA	NCU	NCC	NCG	NGA	NGU	NGC	NGG
Amino Acid	Thr	Ser	Pro	Ala	Ser	Cys	Arg (Orn?)	Gly

(c) Changes in (b) leading to present code as a result of reassignment of tRNAs.

→		NAU		NCU			NCA	
Anticodons	RAU	UAU	CAU	RCU	YCU	RCA	UCA	CCA
Codons ←	YUA	AUA	GUA	YGA	RGA	YGU	AGU	GGU
Amino Acid	Ile	Ile	Met	Ser	Arg	Cys	C.T.	Trp

at this point. The final step is reassignment of several codons, as shown in (c) Table V, so that methionine and tryptophan are added.

To summarize: codon-anticodon pairing at position 1 of the codon may have been ambiguous at an earlier stage in the evolution of the code. The ambiguity disappeared when tRNA evolved to its present specialized state in which the nucleoside at position 40 is often modified, so as to make the pairing specific between position 1 of the codon and position 3 of the anticodon. I therefore postulate that an earlier form of the code could have been as shown in Table V(a). This could reduce the present list of amino acids used in protein synthesis to ten (Jukes, 1973a).

Further support for the possibility of ambiguity in the first position of an early code may be found in the observation that the second base in the codon confers greater specificity on the code than does the first base or, of course, the third base (Woese, 1965). It is therefore logical to attribute more specificity to the second base than to the first base in earlier codes.

The possibility exists that certain amino acids were formerly in the code but have been discarded from protein synthesis as a result of displacement of them from tRNA combination by members of the present 'group of 20'. A mechanism and an example of this was suggested for ornithine. Other amino acids have been found in mixtures obtained by treatment of simpler molecules. These include α-amino-isobutyric acid, α-amino-*n*-butyric acid, β-hydroxyvaline (Miller and Urey, 1959), and β-alanine (Friedman *et al.*, 1972). Whether the ten amino acids in the 'ancestral code' were actually those listed in Table V(a) is open to question in view of the fact that tRNA molecules, as noted above and elsewhere (Jukes and Holmquist, 1972), can 'switch' from one amino acid to another. The main point of the system outlined in Table V(a) is that it would provide for a code in which only ten amino acids would participate.

4. Arginine as an Intruder into Evolution

About one-tenth of the codons (6 out of 61) specify arginine. Yet most proteins contain less than 10% arginine, and in most cases the content is far below this. We have examined the arginine content of 83 proteins each containing 50 or more amino acid residues (Table VI). Some of the 83 are representatives or averages of a group of homologous proteins, such as cytochromes *c* or hemoglobins. Others have no homologs that have been sequenced, so these were treated as the sole representatives of their groups. The arginine and lysine contents of the list of Table VI are compared in Table VII. Lysine is great favored over arginine in proteins, in direct contrast to the comparative numbers of codons for these two amino acids. The disparity is less marked in eukaryotes than in prokaryotes. Perhaps the presence of the ornithine cycle makes mammals more tolerant to arginine by providing a mechanism for its breakdown.

The calculation in the last line of Table VII assumes equal amounts of A, U, C and G in messenger RNA for these proteins, formed by transcription of DNA. The DNA of vertebrates contains about 56% A+T and 44% G+C. If these values represent messenger RNA containing an average of 28% A, 28% U, 22% G and 22% C, then, for vertebrates, the expected percentage of lysine codons in all amino-acid codons would be 3.9% rather than 3.3%, and of arginine codons, 9.1% rather than 9.8%.

A protein with less than 5% arginine contains about half the arginine expected from a random distribution of codons (assuming A=C=G=U) and a protein with more than 6.6% lysine contains more than twice the lysine expected from such a random distribution of codons. *Clearly lysine is greatly favored over arginine as a basic amino acid.* The code evidently contains more arginine codons than are needed, and arginine codons in DNA are genetically reduced in numbers by Darwinian selection (King and Jukes, 1969). In contrast, lysine, which has only two codons, actually exceeds arginine

TABLE VI

Proteins, Eukaryotic	No. of proteins in class	Average			%	
		No. of residues	ARG	LYS	ARG	LYS
Trypsinogen, bovine	1	229	2	15	0.9	6.6
Lactalbumin	3	123	1.3	11.7	1.1	9.5
Ferredoxin – plants	5	96.6	1.2	4.6	1.2	4.8
Leghemoglobin – plants	1	140	2	14	1.4	10.0
Chymotrypsinogen	2	245	4.5	12.5	1.8	5.1
Caseins	2	204	5	12.5	2.5	6.1
β-Lactoglobulin, Bovine AB	1	162	3	15	1.9	9.3
Myoglobin	8	152.9	2.9	19.6	1.9	12.8
Hemoglobin Alpha Chain	10	141.1	3.0	11.6	2.1	8.2
Hemoglobin Beta Chain	10	145.2	3.3	11.3	2.3	7.8
Globin, Chironomus	1	136	3	10	2.2	7.4
Cytochrome *c*	40	107	2.4	15.6	2.2	14.6
Protease Inhibitor – Lima Bean	1	84	2	4	2.4	4.8
Globin, lamprey	2	146	4.5	13	3.1	8.9
Immunoglobuns, light chains:						
Constant, Kappa, human and mouse	2	106	2.5	7.5	2.4	7.1
Constant, Lambda, human and mouse	2	105	1	7	1.0	6.7
Variable Kappa, human and mouse	18	107.8	4.9	3.7	4.6	3.4
Variable, Lambda, human and mouse	10	109	4.0	3.2	3.6	2.9
Immunoglobuns, heavy chain						
Variable, human	6	109.7	5.8	4.2	5.3	3.8
Constant, human	1	336	7	27	2.1	8.0
Constant, rabbit	1	326	13	19	3.9	5.7
Haptoglobin Alpha Chain	1	84	2	8	2.6	10.1
Carbonic anhydrase, human	1	260	7	18	2.7	6.9
Hemerythrin, worm	1	113	3	11	2.7	9.7
Glyceraldehyde 3-PO_4 Dehydrogenase	2	332.5	9.5	27	2.9	8.1
Glutamate Dehydrogenase	1	500	30	32	6.0	6.4
Thyrotropin α and luteinizing hormone α	3	96	3	10	3.1	10.4
Alcohol Dehydrogenase, horse	1	374	12	30	3.2	8.0
Muscle triose PO_4 isomerase	1	248	8	20	3.2	8.1
Chorionic gonadotropin, human α	1	92	3	6	3.3	6.5
Adrenodoxin, bovine	1	118	4	5	3.4	4.2
Thyrotropin β and luteinizing hormone β	2	116.5	10	6.5	8.6	5.6
Carboxypeptidase, bovine	1	307	11	15	3.6	4.9
Prophospholipase A_2, pig	1	130	5	9	3.8	6.9
Ribonuclease, pancreatic	3	125	5	9.6	4.0	7.7
Keratin, High-sulfur protein	2	153.5	6.5	0	4.2	0.0
Trypsin inhibitor, pancreatic	2	56	2.5	3.5	4.5	6.2
Trypsin inhibitor, soybean	1	71	2	5	2.8	7.0
Trypsin inhibitor, ascaris	1	66	3	7	4.5	10.6
Cytochrome B_5	6	87.2	3.6	6.6	4.1	7.8
Bee venom phospholipase A	1	128	6	10	4.7	7.8
Elastase, porcine	1	240	11	2	5.0	1.2
Proinsulin	3	83.6	4	2	4.8	2.4
Lipotropin β and γ	2	90.5	4.5	10	5.0	11.0
Prolactin	2	194	10.5	9	5.6	4.5
Papain	1	212	12	12	5.7	4.7
Mouse nerve growth factor	1	120	7	8	5.8	6.7
Growth Hormone	2	189.5	12	10.5	6.4	5.6

Table VI (Continued)

Proteins, Eukaryotic	No. of proteins in class	Average			%	
		No. of residues	ARG	LYS	ARG	LYS
Parathyroid Hormone	1	84	5	9	6.0	10.7
Neurotoxins, snake venom	11	61.2	4.2	5.3	6.2	8.4
Neurotoxin, scorpion	2	63.5	2.5	5.5	3.9	8.7
Avidin, chicken	1	128	8	9	6.3	7.0
Histone II B_2	1	125	8	20	6.4	16.0
Monkey amyloid protein A	1	76	4	4	6.6	5.3
Trypsin inhibitor, basic, bovine	1	58	6	4	10.3	6.9
β chorionic gonadotropin, human	1	139	4	11	7.9	2.9
Lysozyme – Vertebrates	5	129	11.6	6.2	9.0	4.8
Myelin Membrane Encephalitogenic protein	1	170	18.5	12.5	10.9	7.4
Trypsin Inhibitor – Maize	1	65	8	1	12.4	1.5
Histone III, Calf thymus	1	135	18	13	13.3	9.6
Histone IV	2	102	14.5	10.5	13.7	10.8
	Totals	8909	378.5	617.8	Av. 4.25	6.93
Proteins, prokaryotic						
Rubredoxin	2	52.5	0	3	0.0	5.7
Cytochrome C_2	1	112	0	17	0.0	15.2
Cytochrome C_{551}	3	82	0.3	8.3	0.3	10.2
Cytochrome c_3	1	109	0.5	18.5	0.6	18.5
50S Ribosomal protein A_2, *E. coli*	1	120	1	12	0.8	10.0
Neocarzinostatin *Streptomyces*	1	109	1	0	0.9	0.0
Thioredoxin	1	108	1	10	0.9	9.3
Ribonuclease T_1	1	104	1	1	1.0	1.0
Ferredoxin – clostridial type	5	54.8	0.2	0.8	1.0	1.3
Penicillinase *Staph. aureus*	1	257	4	43	1.6	16.7
Subtilisin	2	274.5	3	10	1.1	3.6
Ferredoxin, *Chromatium*	1	1	81	2	2.5	2.5
Azurin	4	125.8	1.2	12.8	1.2	9.3
Acyl carrier protein	1	77	1	4	1.3	5.2
Coat protein – turnip yellow mosaic virus	1	188	3	7	1.6	3.7
Thermolysin, *Bacillus thermoproteolyticus*	1	316	10	11	3.2	3.5
Nuclease, staphylococcal	1	149	5	23	3.4	15.4
Cytochrome B_{562}	1	110	4	16	3.6	14.5
Tryptophan Synthetase A	1	267	11	13	4.1	4.9
Aspartate transcarbamylase R Chain	1	152	8	10	5.3	6.6
Penicillinase, *B. licheniformis*	1	265	15	24	5.7	9.0
Coat Protein – Tobacco Mosaic Virus	5	157.4	9.6	1.6	6.1	1.3
Alpha lytic protease, myxobact.	1	198	12	2	6.1	1.0
Lysozyme, bacteriophage	2	160.5	12.5	12.5	7.9	7.9
	Totals	3628	106.3	262.5	Av. 2.93	7.24

in most proteins. Lysine ($pK'_3 = 10.5$) confers important basic properties to proteins through its epsilon amino group, but the guanidine group of arginine is far more basic, perhaps excessively so ($pK'_3 = 12.5$). Arginine is formed biochemically from ornithine, ammonia and carbon dioxide via citrulline, or from ornithine and guanidoacetic acid. I suggested (Jukes, 1973b) that ornithine (which is present in peptide linkage in gramicidin) preceded arginine in an earlier genetic code and that the arginine codons were originally assigned to ornithine (Table V(a)). Arginine then appeared as a result of the evolution of the urea cycle, and arginine had a greater affinity than ornithine for the then-existent ornithine tRNAs. Such a phenomenon can occur in enzyme chemistry. Aminopterin (4-amino pteroylglutamic acid) has a much greater affinity than the natural substrate, dihydrofolic acid, for the enzyme dihydrofolic acid reductase.

By replacing ornithine in the aminoacylation of tRNAs, arginine displaced ornithine from protein synthesis. This event increased the sophistication of protein molecules by introducing arginine, a new amino acid with unique properties. It replaced an amino acid (ornithine) which was similar to lysine in the same manner that aspartic acid is similar to glutamic acid, for glutamic acid is homoaspartic acid; lysine is homo-ornithine. The deficiency of ornithine in proteins for needed functions was overcome by increasing the lysine content, accomplished by preferentially selecting lysine codons in evolution. The demand for lysine in the biological synthesis of proteins, incidentally, is so great that lysine deficiency is a major global problem in human nutrition.

TABLE VII

Distribution of Arginine and Lysine in 83 proteins containing 8909 codons

	Eukaryotes	Vertebrates only	Prokaryotes
Arginine			
Less than 5 %	39	29	19
5 %–9.9 %	15	15	5
10 % or more	5	4	0
Lysine			
Less than 3.3 %	6	4	6
3.3 %–6.5 %	15	10	6
6.6 % or more	38	34	12
Arginine: Lysine ratio	0.613	0.613	0.405
for random base sequences	3.0	3.0	3.0

It is perhaps surprising that the codons AGA and AGG are not the third and fourth codons for lysine, which has similar codons. Yeast tRNA for arginine, pairing with AGA, is similar to the yeast tRNA for lysine, pairing with AAA (Table II). These two tRNAs differ in their sequences by only 27% (Jukes and Holmquist, 1972). The average difference between pairs of tRNAs for different amino acids involved in protein synthesis is 49.4% ± 7.0 (Holmquist *et al.*, 1973).

If functional constraints are removed from a protein with low arginine and high lysine content, it should evolve in the direction of higher arginine and lower lysine. I suggested in 1969 that most point mutations in the specificity (S) (variable) regions of immunoglobulins are advantageous, and are rapidly incorporated as evolutionary changes, because 'it is immunologically advantageous to have a large available assortment of different antibodies to cope with various antigenic determinants' while 'those in the constant (C) regions are usually deleterious, thus accounting for the variability of S and the constancy of C sequences,' (Jukes, 1969).

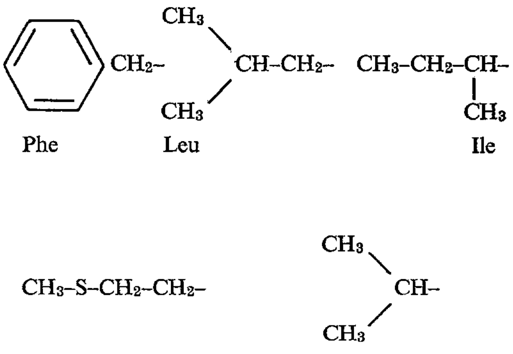
The S and C regions are assumed to have descended from a common ancestor (Hill *et al.*, 1966). The S regions should therefore be higher in arginine and lower in lysine than the C regions as a result of randomly occurring point mutations. Table VI shows that this is indeed the case for both light and heavy human immunoglobulin chains.

5. Aspartic and Glutamic Acids

The similarity of the four codons for the dicarboxylic amino acids is frequently held up as an example of the optimization that exists in the code, on the basis that mutations should be non-injurious because of the similarity of these two amino acids. Aspartic and glutamic acids, by 'internally neutralizing' the basic amino acids arginine, lysine and histidine, are important for formation of zwitterions by proteins, and hence for solubility. However, the similarity may be due to chance, because other pairs of amino acids whose codons are identical except for the third base do not have similar chemical properties. These pairs are asparagine and lysine; histidine and glutamine; cysteine and tryptophan; serine and arginine. I therefore conclude that there is no evidence that the similarity of the Asp and Glu codons is necessarily more than a 'fortunate' accident.

6. The Hydrophobic Amino Acids; Codons with a Middle U

A conspicuous feature of the code is the similarity in chemical properties of the five amino acids having codons with a middle U. These amino acids have the following hydrophobic side chains:



Phe Leu Ile

Met Val

The relationship between codons and properties is often considered to indicate optimization in that the codons for all these amino acids are interconnected by single base changes in the first position. The effect of this is alleged to be that mutations involving the first or third positions of these sixteen codons might tend to be neutral or near-neutral rather than deleterious. Here we have a hint of a process in evolution favoring neutral, non-adaptive changes.

Methionine has a single codon and a special function in protein synthesis as the amino acid that initiates the synthesis of polypeptides, following which it is removed enzymatically when or before the protein is completed. Methionine has other special functions in the interior of polypeptide sequences. We suggest that the elaborate mechanism by which methionine initiates polypeptide synthesis is an evolutionary development that appeared after the code reached its present from, and that valine preceded methionine as the 'initiator' amino acid.

Why do some amino acids have six codons? It was pointed out that so many codons seem unnecessary in the case of arginine. Obviously, however, every possible sequence of three bases in a strand of messenger RNA (and hence in the corresponding DNA sequence) must be translated as an amino acid or serve as a termination signal. There should not be a large number of termination codons in the code, or it will be too difficult to maintain the existence of long polypeptide chains against the frequent occurrence of mutations that produce premature termination, with lethal results. The code therefore cannot contain a high proportion of untranslated codons.

Each codon must contain a minimum of 3 nucleotides to provide for distinguishing between 20 different amino acids.

Most mutations in the third codon position are 'silent' or neutral because they do not change the meaning of the mutated codon. It is also of interest that the six codons for serine are such that each of them can be changed to a codon for the similar amino acid threonine by a single-base change. However, it is not clear why there should be six codons for serine and only two for the much-used glutamic acid. Nor is it clear why there are six codons for leucine and only three for isoleucine. Comparisons of the sequences of tRNAs do not resolve these problems, because the sequences of two tRNAs for *different* amino acids usually differ by about 50% and in most cases, the sequences of two tRNAs for the *same* amino acid in two different organisms also differ by about 50% (Holmquist *et al.*, 1973).

We must answer the question: If at one time the number of amino acids used in protein synthesis was 16 or less, why was the code not 'frozen' as a doublet code? Eigen (1971) answered this on the basis of stability constraints, using data compiled from equilibrium and rate studies with oligonucleotides. He comments:

> The most interesting effect is the preference for the triplet, however, not just for the logically obvious reasons, i.e. the prerequisite for the coding of more than 20 symbols, but rather due to mechanistic coincidences. Codons with less than three digits would be very unstable (at least for A and U). Codons with more than three digits, especially for G and C, become too 'sticky'. The life time of a codon-anticodon pair should not exceed milliseconds so that enzymes with corresponding turnover numbers can adapt optimally. The same type of optimization between stability and rate is always found for enzyme-substrate interactions. Any gain in stability means a lowering of complex dissociation rates; these have to match the turnover numbers in order not to become the rate limiting steps for the turnover.

Gatlin (1972) has found the three-digit codon to be informationally optimal for translation on a theoretical basis. Her calculations show that the entropy H_1 for the the protein 'language', $H_1(P)$, reaches its maximum for a 64-variable source when the codons are triplet sequences.

7. Evolution of tRNA

The sequenced transfer RNAs are a group that includes tRNAs for 16 of the 20 amino acids. Their sequences are sufficiently similar that we can conclude that they have descended from a single ancestral molecule by gene duplication followed by mutations. This conclusion is reinforced by the fact that pairs of tRNAs for the same amino acid in the same organism may differ by as little as 3% or as much as 36% (Holmquist *et al.*, 1973), showing that the process of duplication and evolutionary separation of tRNAs is a continuous and ongoing process. The inference that the tRNAs have a single ancestor leads to the additional conclusion that the original ancestral tRNA could not have accepted more than one amino acid, or contained more than one anticodon. The evolution of tRNA therefore leads us back to the beginnings of the present genetic code, which depends on recognition of adaptor molecules by the codons present in the genetic message. We cannot switch from a 3-letter code to a one containing only two letters per codon without destroying all the accumulated hereditary information, and, in effect, originating life over again.

8. Slippage in DNA Replication

Kornberg *et al.* (1964) discovered that enzymatic synthesis of DNA will take place in the absence of a template to produce repetitive sequences of poly $d(AT)_n$, in the presence of dATP and dTTP as substrates. The process was greatly accelerated by adding oligonucleotides with sequences of 6 to 14 alternating deoxyadenylate and deoxythymidylate residues, such as $d(AT)_6$. The authors postulated that successive stages of replication and slippage resulted in the continuous reiteration of the template and the synthesis of a large dAT co-polymer. This phenomenon was utilized by Khorana *et al.* (1966) in experiments to synthesize repetitive sequences in RNA for solving the problem of the genetic code.

DNA 'families' of identical sequences are recognized as being of comparatively recent evolutionary origin (Britten and Kohne, 1965). Some of these families separate from the main bulk of DNA upon centrifugation and are therefore termed 'satellite' DNA. Satellite DNA of guinea pigs contains a repeating sequence of about six nucleotides (Southern, 1970). Such a molecule would be a possible starting point for coding the synthesis of 'new' polypeptides consisting of repetitions of short oligopeptides that if 'biologically useful', could form functional proteins which could eventually become modified by point mutations. Yčas (1972) has reviewed the existing information on proteins that contain periodic repetitions. Some are remarkably simple, such as a lepidopteran silk fibroin consisting mainly of $(\text{Ala-Gly})_n$. Ycas points out that collagen has possibly evolved from $(\text{Gly-Pro-Pro})_n$, changed by a number of replacements of proline residues by other amino acids.

These repetitive proteins are structural rather than enzymatic in function. They are 'new' in the evolutionary sense, and therefore do not cast light on the composition of earlier genetic codes. We do not know whether enzymes were formed by evolutionary changes in long, repetitive peptide chains that were originally without catalytic activity. Another possibility, which seems more feasible, is that enzymes started as short peptides with weak catalytic activity, and that these short peptides evolved by lengthening to proteins with greater activity, and with structures that enhanced their stability and specificity.

9. Evolution in Transfer RNAs

The genetic code is considered to be universal in terrestrial organisms, and to be 'frozen'. This may be true of the assignments of the codons, but the tRNAs, which function in translation of the code, are evidently still in process of evolution as judged by the presence of duplicate forms of tRNA molecules differing only slightly from each other, such as the two serine tRNAs of yeast (Zachau *et al.*, 1966).

The system that synthesizes polypeptides in cells is extremely complex. It involves ribosomes, which contain large numbers of different proteins as well as RNA. Many enzymes and coenzymes participate in the synthesis of proteins upon ribosomes. The entire process is so elaborate that it appears unlikely to resemble in any respect the primitive systems that produced polypeptides during the period when life was first emerging. A discussion of the origin of ribosomes would be pertinent in an essay on the origin of the code, but, in view of the paucity of information on this subject, I shall not undertake it here.

There are two tRNAs in *Staphylococcus epidermidis* that insert glycine into peptidoglycan molecules, which are units of the bacterial cell wall. These tRNAs differ from the tRNAs that participate in the synthesis of polypeptides upon ribosomes, but the essential differences are only in the absence of T and Ψ from loop IV, and U instead of purine at position 40. The remainder of the molecule, including the regular 'cloverleaf' structure, is quite homologous with the other tRNAs (Holmquist *et al.*, 1973). The synthesis of peptidoglycans, however, may well be an evolutionary development that took place subsequently to the appearance of the tRNA-ribosomal system of polypeptide synthesis.

This evolutionary versatility shown by the tRNAs suggests that they could have undergone an epochal event of rapidly-occurring duplication followed by mutations that brought about numerous changes. These changes could have led to the appearance of a large number of different anticodons and different recognition sites for various amino acids.

If this development took place at the same time as the appearance of the ribosomal RNAs, the fundamental units of a polypeptide-synthesizing system, ribosomes and tRNAs, could have made their appearance together.

Holmquist *et al.* (1973) found that pairwise comparisons, of 43 different tRNAs showed that their primary structures had diverged so far that it is impossible to construct a coherent phylogeny for most of the 43. The average divergence, 49.4%±6.9, for pairs of tRNAs for different amino acids involved in protein synthesis represents an equilibrium between the restraints of natural selection and the flow of point mutations (Holmquist *et al.*, 1973).

The initial appearance of something resembling the present tRNA on the biological scene must have been a portentous event. It could well have taken place suddenly, because, as Eigen (1971) has shown, a 'clover-leaf' structure emerges from a game played with a random sequence of 80 digits composed of A, U, G and C. Each player throws a tetrahedral die, each face of which represents one of the four letters, and tries to approach a double-stranded structure with as many as possible A·U or G·C pairs. This 'game' in evolution would be played by a system that 'tried to resist' hydrolysis by forming base pairs. Once formed, the tRNA molecule could undergo rapid proliferation by gene duplication, and point mutations would soon produce a series of different anticodons, for the anticodon loop would be exposed in single-stranded form by the very formation of the clover-leaf structure. Such a 'family' of tRNAs could furnish enough anticodons to pair with all the codons, especially if codon-anticodon pairing were only partially specific, as is the case in the third position of codons. If pairing were relaxed in the first as well as in the third position, a pattern could exist as shown in Table V(a). The pattern of amino acid binding by tRNA might develop more slowly than the changes in anticodons; at first, only one or two amino acids might be bound by the primordial tRNAs. Advantages would soon accrue to the system that built up the number of amino acids that were bound by tRNAs and thus were brought into protein synthesis.

The argument that optimization of the code is shown by the nature of amino acid

TABLE VIII

Amino acid interchanges that can occur as the result of single-base changes in the coding triplets

Amino acid	Possible interchanges[a]
Ala	Asp Glu Gly Pro Ser Thr Val
Arg	Cys Gln Gly His Ile Leu Lys Met Pro Ser Thr Trp
Asn	Asp His Ile Lys Ser Thr Tyr
Asp	Glu Gly His Tyr Val
Cys	Gly Phe Ser Tyr Trp
Gln	Glu His Leu Lys Pro
Glu	Gly Lys Val
Gly	Ser Trp Val
His	Leu Pro Tyr
Ile	Leu Lys Met Phe Ser Thr Val
Leu	Met Phe Pro Ser Trp Val
Lys	Thr Met
Met	Thr Val
Phe	Ser Tyr Val
Pro	Ser Thr
Ser	Thr Trp Tyr

[a] The underlined examples have been reported to occur in mutations.

replacements resulting from single-base changes is unconvincing. The 75 possible amino acid interchanges resulting from single-base changes are in Table VIII. It is obvious that many of them represent interchanges between amino acids of widely differing properties. We conclude that the only optimization shown by the code is the fact that many of the changes in the third base of codons do not produce changes in amino acids. This feature may be an incidental result of the spatial nature of codon-anticodon pairing rather than an 'evolutionary optimization'.

10. Origin of a Translation System

The proposal that the genetic code originated by some kind of loose affinity between amino acids and the bases of nucleic acids is an obvious one. Indeed, it was the basis of the first suggestions for a genetic coding system made by Dounce (1952) and Gamow (1954). These suggestions were discarded when they were put to flight by Crick's 'adaptor hypothesis' in 1955. However, although the above proposal is obvious, it runs into very serious practical difficulties. Some of these were discussed by Razska and Mandel (1972). It is fairly easy to show that some amino acid, such as phenylalanine, has a slight preferential affinity for binding weakly to some base, such as uracil, under specialized conditions, but nothing has emerged from these experiments that indicates a special attraction of each individual codon or (just as logically) each anticodon, for the appropriate member of the list of 20 amino acids. This defect is usually countered by saying that even a slight preference in binding, spread over hundreds of millions of years of chemical trial and error, would result in the emergence of a coherent code. Serious objections can be made to this rationalization, because the essence of a successful coding system is *fidelity*, and the errors arising from a pairing mechanism containing a high degree of indefiniteness or ambiguity would be so great as to prevent the emergence of the process of heredity. This objection was raised and discussed by Eigen (1971).

The protagonists of the 'weak binding' theory for the origin of the code include Woese (1967) whose viewpoint was summarized by him in 1967 as follows:

> ...amino acid-nucleic acid 'recognition' interactions (or their equivalent), ... being very weak interactions... could not have given rise directly to the genetic code as it now exists. Instead they must have played the role of constraints operating on the evolutionary process, in this way gradually shaping the form of the code. Since the interactions are weak and therefore cannot manifest an all-or-none sort of specificity with regard to amino acid-codon pairings, it is reasonable to expect that they cannot align amino acids along a nucleic acid template directly, and so this 'recognition' role has been filled by the evolution of an 'intermediary' system, the tRNA's and activating enzymes, that recognize with very high accuracy both an amino acid and its codons.

An objection to this line of reasoning is that it tends to invoke the existence of interactions that are so weak that they cannot be detected except under specialized and artifical experimental conditions. I feel that such an explanation may rely on a desire to postulate non-existent phenomena because we feel intuitively that they ought to have existed since we cannot think of any other way that the genetic coding of proteins might have started. There are infinitely large numbers of ways in which long sequences of variables can be arranged. The task of natural selection working alone to find useful members of an infinitely large set would be impossible. Therefore, there is a tendency to hope that there must have been an orderliness in aligning amino acid with nucleic acids in some primitive system.

A strong objection to the 'stereochemical fit' concept is that the two sets of codons for serine, UCN and AGY, are so dissimilar. Various authors have proposed that there is a relationship between the second base of codons and the chemical properties of the amino acid, and that this relationship has governed the evolution of the code. Such a relationship, in my opinion, is perceptible only for the codons with U in the middle position. The other three 'groups' (Table I), do not share common properties. For example, although serine and threonine are similar, they do not resemble proline; neither are cysteine, tryptophan and arginine similar to each other. Moreover, serine codons are found in two different classes. The 'clustering' of the hydrophobic amino acids could have come about by the phenomenon postulated in Table V, in which all 16 NUN codons are shown as having evolved through three stages of increasingly precise codon-anticodon pairing. The fact that five amino acids with hydrophobic side-chains all have codons with a middle U may have been shaped by evolution, or it may be a coincidence.

TABLE IX

Relative abundance of certain elements in the Earth's crust and in sea-water (After Mason, 1952; Sverdrup *et al.*, 1942; Bowen, 1966)

Element	Presence in Earth's crust ppm	Sea-water pp 10^9
Fe	50000	2 to 20
Mn	1000	1 to 10
Cr	200	0.05
Zn	132	5 to 14
Ni	80	0.1 to 5.4
Cu	70	1 to 9
Co	23	0.1 to 0.27
Mo	15	0.3 to 10
Se	0.09	0.09 to 4

Mention should be made of the romantic suggestion by Crick and Orgel (1973) that the genetic code could be of pan-spermatic, extra-terrestrial origin; coming from an extra-galactic source via space-ship. In their proposal, Crick and Orgel go to the length of inferring that the beings who sent us the code dwell in an environment that is rich in molybdenum. Their argument is that terrestrial organisms, in contrast to the Earth's crust, are higher in molybdenum than in nickel. However, the proportions of minerals in animals are thought to resemble the composition of these found in sea-water rather than that of the geosphere. The six trace elements that are components of identified enzyme systems are iron, zinc, copper, manganese, molybdenum and selenium. A seventh, cobalt, is not utilized as an inorganic element; animals (but not green plants) use it as a component of vitamin B_{12}, which they obtain from microorganisms. It is interesting to compare the six in terms of their abundance in the Earth's crust and in sea-water (Table VI). They show a tremendous range in their abundance in the earth's crust, extending between six and seven orders of magnitude. However, their concentrations in sea-water, in which life is thought to have originated, are remarkably uniform. The average values for each of the six 'enzymatic' trace elements extend from 1 to 11 parts per 10^9. Note that chromium occurs below this range, and so does cobalt. Molybdenum occurs in sea-water at higher concentrations than nickel or chromium. It is unnecessary to postulate that molybdenum stars might have served as a jumping-off point for pan-spermatic organisms that 'infected' the Earth. In any case, the mineral content of protoplasm evidently is greatly influenced by natural selection, judging from anomalies such as the iodine content of the thyroid gland, the vanadium content of tunicates, etc.

11. Summary

(1) Suggestions are made for possible pathways of evolution of the genetic code, assuming that the present code was preceded by codes that contained fewer amino acids. It is recognized, however, that it is possible that earlier codes contained amino acids that are not currently used in protein synthesis.

(2) As an example of the latter phenomenon, it is suggested that ornithine preceded arginine in the code. Support for such a suggestion is found in the observation that the average arginine content of proteins is only about half the value expected from the fact that arginine has six codons. It is proposed that the introduction of arginine led to a diminished use of the codons CGN and AGR, and an evolutionary selection of more lysine resulted than would be expected from the fact that it has only two codons.

(3) Pairing between G and U (or U and G), in the first position of codons and the third position of anticodons could reduce the number of amino acids involved in protein synthesis to ten. It is suggested that this formed the basis of an earlier code.

(4) The introduction of tRNA into protein synthesis may have been a cataclysmic and comparatively sudden event, since duplication of tRNA takes place readily, and point mutations could rapidly differentiate members of the family of duplicates from each other.

Acknowledgements

This work was supported by NASA Grant NGR 05-003-460 to the University of California, Berkeley.

I wish to dedicate this essay to the memory of Prof. J. B. S. Haldane, who asked me to write a manuscript for him during the meeting on 'The Origins of Prebiological Systems and of their Molecular Matrices' at Wakulla Springs, Florida, November 1963.

References

Bowen, H. J. M.: 1966, in *Trace Elements in Biochemistry*, Academic Press, London and N.Y.

Britten, R. J. and Kohne, D. E.: 1965–1966, *Carnegie Institute*, Washington, D.C. 20015, Year Book, pp. 78–106.

Crick, F. H. C.: 1966, *J. Mol. Biol.* **19**, 548.

Crick, F. H. C. and Orgel, L.: 1973, *Icarus* **19**, 341.

Dounce, A. L.: 1952, *Enzymologia* **15**, 251.

Eigen, M.: 1971, *Naturwissenschaften* **58**, 465.

Friedman, N., Haverland, W. J., and Miller, S. L.: 1972, in R. Buvet and C. Ponnamperuma (eds.), *Molecular Evolution* **1**, 123, North-Holland, American Elsevier Publishing Co., Inc., New York.

Gamow, G.: 1954, *Nature* **173**, 318.

Gatlin, L. L.: 1972, *Information Theory and the Living System*, Columbia University Press, New York, pp. 1–210.

Hill, R. L., Delaney, R., Fellows, R. E., Jr., and Lebovitz, H. E.: 1966, *Proc. Nat. Acad. Sci. U.S.A.* **56**, 1762.

Holmquist, T., Jukes, T. H., and Pangburn, S.: 1973, *J. Mol. Biol.* **78**, 91.

Jukes, T. H.: 1966, *Molecules and Evolution*, Columbia University Press, New York and London, pp. 1–285.

Jukes, T. H.: 1969, *Biochemical Genetics* **3**, 109.

Jukes, T. H.: 1973a, *Nature* **246**, 22.

Jukes, T. H.: 1973b, *Biochem. Biophys. Res. Commun.* **53**, 709.

Jukes, T. H. and Holmquist, R.: 1972, *Biochem. Biophys. Res. Commun.* **49**, 212.

King, J. L. and Jukes, T. H.: 1969, *Science* **164**, 788.

Khorana, H. G., Buchi, H., Ghosh, H., Gupta, N., Jacob, T. M., Kossel, H., Morgan, R., Narang, S. A., Ohtsuka, E., and Wells, R. D.: 1966, *Cold Spring Harb. Symp. Quant. Biol.* **31**, 39.

Kornberg, A., Bertsch, L. L., Jackson, J. F., and Khorana, H. G.: 1964, *Proc. Natl. Acad. Sci. U.S.A.* **51**, 315.

Mason, B.: 1952, *Principles of Geochemistry*, John Wiley and Sons.

Miller, S. L. and Urey, H. C.: 1959, *Science* **130**, 245.

Murao, K., Saneyoshi, F., Harada, F., and Nishimura, S.: 1970, *Biochem. Biophys. Res. Commun.* **38**, 657.

Nishimura, S.: 1972, *Progr. Nucl. Acid. Res. Mol. Biol.* **12**, 49.

Raszka, M. and Mandel, M.: 1972, *J. Mol. Evol.* **2**, 38.

Southern, E. M.: 1970, *Nature* **227**, 794.

Stewart, T. S., Roberts, R. J., and Strominger, J. L.: 1971, *Nature* **230**, 36.

Sverdrup, L., Johnson, W., and Fleming, J.: 1942, *The Oceans* Prentice Hall.

Woese, C. R.: 1965, *Proc. Nat. Acad. Sci. U.S.A.* **54**, 1546.

Woese, C. R.: 1967, *Progr. Nucl. Acid Res. Mol. Biol.* **7**, 107.

Ycas, M.: 1972, *J. Mol. Evol.* **2**, 17.

Zachau, H. G., Dütting, D., and Feldmann, H.: 1966, *Z. Physiol. Chem.* **347**, 212.