ENCODING AND DECODING OF GENETIC INFORMATION[1]

P. C. HUANG and R. C. HUANG[2]

(Received Jan. 30, 1963)

Introduction

The concept that continuity of organisms resides with a discrete genetic mechanism is now probably irrefutable. The genetic material is characterized by its duplexity, expressivity, mutability, multiplicity, species specificity, helicity, stability, and universality. While the neoclassic concept of one-gene-one-enzyme of Beadle and Tatum (1941) correlates biochemically and physiologically to the Mendelian gene as Morgan defined it (Morgan, 1928) and the biological functions, evidence that DNA (deoxyribonucleic acid) is the major genetic material has been overwhelming as for example the results from transformation (see Hotchkiss, 1957) and transduction (see Hartman, 1957). Then Wilkin, Watson and Crick (see Watson and Crick, 1953) have provided a physical-chemical basis of the DNA molecule. Equally important are the findings of Pauling *et al.* (see Ingram, 1957) that several forms of abnormal human hemoglobin differ from normal only by a single amino acid replacement and each of these changes could be attributed to a single gene defect. Such findings open up the possibility of understanding the relations between chemical structures of DNA and proteins, and most enthusiastically the possible correspondence between the base sequence of DNA (and RNA) and the arrangement of amino acids of the protein (and enzyme). Today, the physical structure of a gene can be resolved to the nucleotide level. A gene can also be viewed as cistron (functional unit), recon (recombinational unit), or muton (mutational unit) (see Benzer, 1957) and recognized at the regulatory, operational as well as structural level (Jacob and Monod, 1961). The amino acid sequence is now known in many proteins (see e.g. Eck, 1962).

The role of RNA (ribonucleic acid) in the transfer of genetic information is getting to be understood. It is now believed that three types of RNAs are involved. One type, called messenger RNA (mRNA), is formed on DNA using the latter as a template and reflects the specificity of DNA in terms of its

(1) Dedicated to Professor Dr. Liang-fang Chao, Academician, Academia Sinica on the seventieth anniversary of his birthday.

(2) Division of Biology, California Institute of Technology, Pasadena, California, U. S. A.

complementary base sequences (Volkin and Astrachan, 1956; Bonner *et al.*, 1961; Spiegelman, 1961; Hurwitz *et al.*, 1962; Chamberlin and Berg, 1962). This mRNA has been shown to reflect the genetic code by aligning specific amino acids of polypeptides in a sequenced order at the second type of RNA, ribosomal RNA (Brenner, 1961). The third type, termed transfer RNA (tRNA) functions as an adaptor for specific amino acids during the code sequence recognition process. The exact mechanism involved in this process is yet to be clarified, but it has been shown recently that this tRNA is complementary to mRNA in base composition (Weisbleum *et al.*, 1962).

It has also been shown that while a native DNA molecule except a few such as the coliphage ϕX174 (Sinsheimer, 1959) is double stranded, only one of them is used in priming for the mRNA (Champe and Benzer, 1962). The ability to promote incorporation of amino acids into polypeptides is similar to both native and synthetic polyribonucleotides (Wood and Berg, 1962). *In vitro,* as long as mRNA is abundant and active, a controlled synthesis of a specific protein can be observed regardless of the origin of ribosomal RNA. However, for the messenger to be continually synthesized *in vivo,* the integral presence of genetically active DNA is required.

It should be pointed out that there are certain cellular phenomena, such as catabolite repression (glucose effect), feedback control, allosteric repression, enzyme induction, that may only be explained in part by direct correlation with the primary nucleotide sequence of either DNA or RNA (see Cold Spring Harbor Symp. Quant. Biol., 1961). The DNA-dependent RNA synthesis, for instance, may be inhibited by histone, suggesting that histone may act as a suppressor of genetic activity (Huang and Bonner, 1962). Also proteins such as antibiotics may be formed through a simple repeating mechanism which requires no direct control by a genic code (Ito and Strominger, 1960). DNA probably does however exert an ultimate and indirect control over all these other processes and phenomena. The specificity and regulation of gene expression have been extensively reviewed (see e.g. Riley and Pardee, 1962) and shall not be discussed here.

With all these reservations, the current belief is then that the genetic information is stored in and transmitted by DNA, but is transcribed into, and expressed by the language of RNA for the synthesis of functional proteins. Thus, one of the most interesting problems of molecular biology today is to understand the relationships among DNA, RNA, and proteins; or in other words the genetic material, the messenger for the genes, and the functional expression of the genes. This paper will consider some recent advances in this area, particularly those related to the problem of encoding and decoding of the genetic information.

Theories and Methodology in the Formulation of Coding Models

The so called coding model calls for a correlation between the composition of the nucleic acids and proteins. Particularly concerned are the four nucleotide bases adenine (A), guanine (G), cytosine (C), and thymine (T) in DNA or A, G, C, and uracil (U) in RNA and the 20 most commonly found amino acids as abbreviated below:

ala	alanine	lys	lysine
arg	arginine	leu	leucine
asp	aspartic acid	met	methionine
asn	asparagine	phe	phenylalanine
cys	cystine, cysteine	pro	proline
glu	glutamic acid	thr	threonine
gln	glutamine	try	tryptophane
gly	glycine	tyr	tyrosine
his	histidine	val	valine
ilu	isoleucine	ser	serine

The majority of the direct approaches to the problem of encoding and decoding of genetic information have appeared only rather recently.* These include approaches from theoretical, biometrical, biochemical and biological points of view. These attempts have been ingenious; although many have been shown to be erroneous. One may refer to Gamov *et al.* (1956), Ycas (1962), and Crick (1962) for general background of the coding hypothesis.

a. *Theoretical approach*

Historically. Dounce (1952) should be given credit for originating the coding idea. He suggested that there should be a dictionary for encoding and decoding all the genetic information. Gamov (1954) (see Gamov *et al.*, 1956) however, was the first to present a detailed, fundamental consideration of the problem. Their main assumption was that a correlation exists between the amino acid sequence of proteins and the pre-existing unique sequence of the nucleotide bases in DNA. This school of thought has dominated all the subsequent theses on the subject and has received the most experimental support.

Two different theoretical considorations of the problem have been presented by Rosen (1960, 1961a, 1961b) and Pattee (1961). Rosen (loc. cit.) suggested that correlations between amino acid compositions of polypeptides and nucleotide compositions of DNA may be unsuccessful due to the existence of anomalous codes. An anomalous code is not completely determined by any definite set of initial conditions. Based on the Hilbert space operation, he assumed that the genetic information is carried by a family of numerical observables of a specific

* The survey of the literature pertaining to this topic was concluded on December 31, 1962.

microphysical system; each of which is sufficient to carry the information. The quantum-theoretic notion of such a microphysical system embodies the idea that what one actually observes in nature is not the system itself but a state in which the observables comprising the system assume more or less definite numerical values. There are many states of the genetic observables which cannot be decoded by the biological system concerned. In this case, there cannot exist a coding model for the transmission of genetic information. An interesting corollary of this theory is that the genetic observables possessing degenerate eigenvalues may be the reason for both multiple allelism and pseudoallelism. The eigenvalues of their states may lie so close together that the probability of them being separated by an observing apparatus with limited resolution power, e.g. recombination and complementation tests, is significantly less than unity.

Josse, Kaiser, and Kornberg (1961) reported that DNA dinucleotide frequency in various organisms deviate only little, although systemically, from the frequency expected from a random sequence of the bases. Pattee (1961) proposed that simple computation with feedback mechanism can assemble, elaborate, and repeat any ordered sequence without requiring the pre-existing ones. The genetic mechanism today is thus the result of the evolution of naturally occurring ordered macromolecular sequences as Freese (1962) predicted on a pure mathematical ground but not of a chance origin. Pattee noticed that there is a long and short range intramolecular periodicity in TMV and fibrous proteins. He argued that any regularities or periodicities in sequences represent redundant information and need not be determined individually by genetic information. The fact that the addition of amino acids in a growing sequence of protein is not positionally random but steadily sequential from one terminal end indicated to him that at least some transfer of information from the growing sequence is necessary. He pointed out that all amino acids must be present for both RNA and protein synthesis; therefore there must exist an interdependent RNA-protein sequential growth process. But we know now that not all amino acids are required to be present for the synthesis of RNA *in vitro*. Refuting this basis of reasoning, however, does not imply invalidity of all of his speculations. The distribution of amino acids in some proteins does suggest some degree of sequential computation starting from the N terminal end of the protein chain. Eck (1962) argued that the amino acid distribution in a protein is probably nonrandom, but subject to systematic constraints. For instance, if cysteine is replaced by another amino acid the newly formed sulfur cross linkage, if any, may be so changed that it is 'lethal' to the molecule. Thus in proteins, a high order of limited variation of amino acid is permitted in the sequence structure. Earlier, Ledley (1955) proposed a code on the basis of symbolic logic, by directly

assigning the amino acids to the nucleotide bases and computing the frequency of likelihood. However, the final construction of such a code is possible only if electronic computers can be efficiently used.

Most of the research on coding, however, is based on the presumption that genetic information is encoded by a simple dictionary in a pre-existing linear order of the consecutive nucleotide bases of DNA, thus of RNA, and decoded in the form of a linear sequence of amino acids in the functional proteins. Most of the coding models proposed by this school of thought are characterized by their linearity, time-independency, freedom from intersymbol restrictions and from the growing configuration of the molecule, and by their freedom from an information feedback mechanism. Crick (1962) termed each code word a "codon."

Table 1. *Gamov's code—1954*

1. AAA	11. GGA, GAG, AGG
2. CCC	12. GGC, GCG, CGG
3. GGG	13. GGT, GTG, TGG
4. TTT	14. TTA, TAT, ATT
5. AAC, ACA, CAA	15. TTC, TCT, CTT
6. AAG, AGA, GAA	16. TTG, TGT, GTT
7. AAT, ATA, TAA	17. ACG, AGC, CAG, CGA, GAC, GCA
8. CCA, CAC, ACC	18. ACT, ATC, CAT, CTA, TAC, TCA
9. CCG, CGC, GCC	19. AGT, ATG, GAT, GTA, TAG, TGA
10. CCT, CTC, TCC	20. CGT, CTG, GCT, GTC, TCG, TGC

Uniform triplets, overlapping, all meaningful.

The first code proposed by Gamov *et al.* consists of exactly 20 sets of triplets. There are 4 kinds of bases and 20 or so kinds of amino acids. Taking 3 bases (triplet) at a time, gives 64 combinations. If the order of the bases in a triplet can be considered irrelevant to its meaning, there are exactly 20 sets of triplets; each seemingly corresponding to one amino acid. This code is degenerate, i.e. having more than 1 code word standing for the same amino acid and totally decipherable (see Table 1). Theoretically, a totally decipherable code is one in which every sequence of symbols has an interpretable message, i.e. there is no nonsense in the text. In nature, it is known that there are regions of the genetic map which are apparently nonsense, suggesting that such a property may not be valid. Another argument against such a code is that if enough restrictions are applied to a set of functions, one is always able to group them into any number of subsets. The fact that Gemov's triplets numbered 20 is fortuitous. Another difficulty with this code is that the coding sequence is overlapping. Thus a sequence of ..ATTUGA.. could be broken down into

..A, .AT, ATT, TTU, TUG, UGA, GA., and A.. This would imply a strong constraint on the type of transition which can occur from one amino acid to the next in a sequence. Brenner (1957) later showed that too many different transitions occur in nature and therefore this type of overlapping code cannot be involved.

Crick *et al.* in 1957 proposed a triplet coding system different from that of Gamov *et al.* (loc. cit.) suggesting that the code is not degenerate, not overlapping, but is commafree and contains nonsense. Commafree means an imaginative comma is placed between the adjacent code letters breaking the sequence into independent code words. Thus if AGA and CGC were two code words situated next to each other in a sequence the overlapping combinations such as GAC and ACG would not be meaningful. Ironically, out of the 64 possible combinations of triplets only 20 can be chosen to form such a unique set (see Table 2). This code, unique as it is, excludes code words, such as UUU and CCC, which have been shown to have a specific priming function on amino acid incorporation (see later). Hersch (1962), studying mutants of TMV suggested that the code is commaless.

Table 2. *Crick's code—1956*

1. ACA	6. CGA	11. ATG	16. CTT
2. ACC	7. CGC	12. ATT	17. GTA
3. AGA	8. CGG	13. CTA	18. GTC
4. AGC	9. ATA	14. CTC	19. GTG
5. AGG	10. ATC	15. CTG	20. GTT

Uniform triplets, commafree, non-degenerate, some nonsense.

Frendenthal was also in favor of the triplet coding system. He showed (1958) that 5 and only 5 basic types of codes (Table 3) could be constructed which would satisfy the commaless condition and contain 20 unique combinations. In each group with a given middle letter, all recombinations of the listed first and third letters are sense words. It is hard to conceive *a priori* that these five unspecified groups should ever be classified as such, although it may be mathematically feasible. There have been suggestions concerning the fixation of certain letter(s) in a code. Petruska (1962) for example has tried to specify two of the three positions in a triplet, varying the other one to fit the then published amino acid replacement data. For instance, if a code is to begin with UG, it may end with either U, C or A, G. In other words, in the third position, the two pyrimidine are essentially equivalent and so are the two purines (see Table 6). Their assignment was very similar to the doublet code of Roberts (1962a, b) and both are in fair agreement with those independently

Table 3. *Frendenthal's code (1958) as arranged by Levinthal (1959). Five possible types of commaless codes using 4 different letters taken 3 at a time*

I	II	III	VI	V
A A B D B C C D	A A B D B C C D	A A B D B C C D	A A B D B C C	A A B D B C D
A A C B B C	A A B C B C	A A C B C	A A B C B D C	A A B C B D C
A A B B	A B B ——— B A A	A A B C B	A B A B	A B A ——— B A A

In each group with a given middle letter, all combinations of the listed first and third letters are *sense* words.

suggested by experimentalists. Smith (1962) also suggests that U occupies the same, but unknown, position in a triplet code for 16 amino acids but varies in position for four other amino acids.

Golomb (1962) extended his earlier work with Delbrück and Welch (1958) and suggested a sextuplet code, with 24 unique combinations (Table 4). It has a special feature of error detecting and correcting as it requires two simul-

Table 4. *Golomb's code—1960*

1. TTTTTG	7. GTGGCC	13. AAAAAC	19. CACCGG
2. GCACTA	8. TTCAGC	14. CGTGAT	20. AAGTCG
3. GGATGT	9. TGGCAA	15. CCTACA	21. ACCGTT
4. TACTCC	10. TCAGAG	16. ATGAGG	22. AGTCTC
5. GATGGA	11. GGCAGG	17. CTACCT	23. CCGTGC
6. GCTCAT	12. TAGATT	18. CGAGTA	24. ATCTAA

Uniform sextuplet, commafree, non-degenerate, some nonsense.

taneous mutations to effect a change in the coding. The code is nondegenerate and is commaless. The assignments, however, have not received experimental support. The code seems to be redundant and inefficient. Later (1962) he returned to favor triplets and proposed 2 sets of codes, thesis and antithesis, as shown in Table 5 (personal communication). The thesis code is based on the assumption that among the double strands of DNA, only one encodes information which is to be transcribed and the other is redundant and probably is involved in the replication of DNA itself. Therefore, for a particular base pair,

Table 5. *Golomb's thesis and antithesis codes (1961)*

Thesis		Antithesis	Thesis		Antithesis
UUU	AAA	AAA	GGU, UGG	ACC, CCA	UCC, CUC, CCU
UUA, AUU	UAA, AAU	CCC	GUG	CAC	UUC, UCU, CUU
UAU	AUA	GGG	ACG, GCA	CGU, UGC	UGG, GUG, GGU
CCC	GGG	UUU	CAG, GAC	CUG, GUC	UUG, UGU, GUU
CCG, GCC	CGG, GGC	AAU, AUA, UAA	AGC, CGA	GCU, UCG	ACG, CGA, GAC
CGC	GCG	AUU, UAU, UUA	CAU, UAC	AUG, GUA	AGC, GCA, CAG
UUC, CUU	GAA, AAG	CCG, GCC, CGC	CAU, UCA	AGU, UGA	ACU, CUA, UAC
UCU	AGA	CGG, GCG, GGC	CUA, AUC	UAG, GAU	AUC, UCA, CAU
CCU, UCC	AGG, GGA	AAC, ACA, CAA			AGU, GUA, UAG
CUC	GAG	ACC, CAC, CCA			AUG, UGA, GAU
UUG, GUU	CAA, AAC	AAG, AGA, GAA			CGU, GUC, UCG
UGU	ACA	AGG, GAG, GGA			CUG, UGC, GCU

See text for explanation

say G-C, if G is meaningful C must be nonsensical. The advantages of eliminating the complementary triplets may lie in freeing the RNA from being interferred in its function by intra- and intermolecular hydrogen bonds. The weak point of the thesis code is apparently its stringent restriction and inefficiency. Here half of the possible code words are nonsense, and no experimental results have supported this view. It may allow code assignments for the presence of the less frequent amino acids such as hydroxyproline and hydroxylysine. However, the antithesis code with 24 sets of code words has the advantage of being flexible. It allows both the requirements for nonsense and degeneracy. The antithesis code is constructed on the basis that all permutations of the letters within a word specify the same amino acid. For instance AAC, ACA, and CAA would carry the same code. Golomb specifies, however, that this antithesis code does not imply overlapping as the relationship of the permutating combinations seemingly suggest. Both sets of codes can be subject to tests. Exactly the same type of code is proposed by Ageno (1962) very recently but indepen-

dently. Neither Golomb nor Ageno was at the time of preparing of their theses aware of the fact that the same consideration had previously been considered and treated mathematically by Chavcharidze in 1958. Chavcharidze calculated further that the entropy of various codes is different. While he was uncertain about the coding ratio, he postulated that either a triplet such as $\overline{AGC}$ or a duplex-triplet such as $\frac{AGC}{TCG}$ (a triplet together with its counterparts) could be meaningful. Indeed, even with the recent knowledge on the priming ability of DNA, these alternatives are still hard to be distinguished.

A different idea concerning the coding ratio was presented by Sinsheimer (1959). He suggested that each code consists of only two bases according to the presence of a 6 amino or a 6 keto group on the nucleotides. Such a hypothesis was made to explain the near equivalence of 6 amino and 6 keto nucleotides in the total RNA of various organisms. On the basis that while the purine and pyrimidine ratios of RNA vary little from organism to organism (A+U/G+C=1.03 to 1.45), G/A hence C/T of DNA varies greatly (0.45 to 2.7), he reasoned that either there is a 2-letter code or there is not a universal code relating DNA to RNA or DNA to protein. The second argument is probably invalid because the kinds of amino acids constituting proteins in all terrestrial organisms have been confined to about 20. Since this time, it has become known that only some RNAs which code for amino acids have a homology in base equivalence to DNA. However, the idea that a dinucleotide may be sufficient as an encoding unit has been upheld by Roberts (1962a, b).

Knowing the composition of amino acid composition of *E. coli* protein and assuming the codes assigned earlier by Nirenberg and Ochoa (see later) Roberts was able to calculate the nucleotide composition of a hypothetical mRNA for *E. coli*. This mRNA had an extremely high U content (45%) compared to natural RNA. When U, which is common to all triplet codes suggested, is discarded in the calculation, the composition of the hypothetical messenger RNA is comparable to that of the 50s ribosomal RNA when allowing for the priming efficiency of the messenger RNA. Robert's doublet assignments are shown in Table 6. This code model contains no nonsense, all 4^2 combinations are used. He later (1962b) showed that the frequency of purine-purine, G-C and amino-keto pairings are higher in nature. It should be noted that our supposition has been that the template for protein synthesis lies with the DNA-complementing mRNA. If Roberts' correlation were valid, it would suggest that the ribosomal RNA could also serve as a template. Roberts pointed out the life time of the ribosomal RNA is long enough to allow it to serve as template for 20–40 polypeptide strand and there is so far no kinetic evidence against this possibility.

Table 6. *Summarized duplet and triplet code assignments*

Sequence unspecified except underlined with——

	Ochoa (Dec. 19, 1962, VIII) in press	Nirenberg (1962) in press	Roberts (1962)	Petruska (1962)	Woese (1962)	Correlation with G+C of DNA as predicted by Sueoka (1961)
ala	CUG, CAG, CCG	CCG	..GC	GCU, GCC, GCA, GCG	CUG, CGG	+ +
arg	GUC, GAA, GCC	CGC	..CG	CGU, CGC, AGA, AGG	UAG, UCG	+ +
asn	UAA, CUA, CAA	UAC	..AA	CCA, CCG, ACU, ACC	UAA, UCA	} 0 to −
asp	GUA, GCA	- - -	..AG	AGU, AGC	AUG, AGG	
cys	GUU	UUG, UGG	..UG	GUU, GUC	UUG, UGG	?
glu	AUG, AAG	ACA, AGA, AGU	..GA	GAU. GAC, GAG, GAG	GUA, GGA	} 0 to −
gln	AGG, AAC	- - -	..GA	CGU, CGC	GAA, GCA	
gly	GUG, GAG, GCG	UGG	..GG	GGU, GGA, GGC, GGG	GUG, GGG	+
his	AUC, ACC	ACC	..CA	ACA, ACG	CUA, CGA	−
ilu	UUA, AAU	UUA	..AU	UAU, UAC, UAA, UAG	AUU, AUC, AGU, AGC	− −
leu	UAU, UUC, UGU	GUU, CUU, (UUU)	..CU	UCU, UCC, UUA, UUG	CUU, CUC, CGU, CGC	0
lys	AUA, AAA	AAA, AAC, AAG, AAU	..AA	AAU, AAC, AAG, AAA	AUA, AGA	− −
met	UGA	UGA	..AG	UGA, UGG	GAU, GCU, GAC, GCC	− or − −
phe	UUU	UUU	..UU	UUU, UUC	UUU, UUC, UGU, UGC	− −
pro	CUC, CCC, CAC	CCC, CCU, CCA, CCG	..CC	CCU, CCC	CAU, CCU, CAC, CCC	+
ser	CUU, ACG	UCG, UCU	..UC	CUU, CUC, CUA, CUG, UCA, UCG	UAU, UCU, UAC, UCC	−
thr	UCA, ACA, CGC	GAC, CAA	..AC	CAU, CAC, CAA, CAG	AAU, ACU, AAC, ACC	0
try	UGG	UGG	..GG	GUA, GUG	- - -	?
tyr	AUU	UAU	..UA	AUU, AUC, AUA, AUG	UUA, UGA	− −
val	UUG	UGU	..GU	UGU, UGC	GUU, GGU, GUC, GGC	0

b. *Biometric approaches*

There are two types of studies involving analysis of published results with biometric techniques. One correlates amino acid differences between related proteins of different species. The other correlates amino acid changes in a specific protein resulting from single induced mutational event.

Ycas (1960, 1961) from correlation studies between viral protein and nucleic acids and from amino acid replacement analyses of various proteins suggested a coding ratio of 1. He advocated that one nucleotide is sufficient to determine one amino acid. Amino acids can be assigned to nucleotides in such a manner that the mole fraction of each group of amino acids in a protein is equal to the mole fraction of the corresponding nucleotide in RNA. Thus:

A—glu, gln, gly, leu, phe, try
U—asp, asn, ilu, his, ser
G—arg, ala, tyr, val
C—cys, lys, met, pro, thr

He argued against the supposition that neighboring nucleotides provide additional information. He implied that viral RNA possesses only part of the information required to specify a protein. Each nucleotide would limit, but not determine, the choice of a residue at any given position. Additional information is contributed by some other structure in the host. He suggested that TMV RNA functions analogously to RNA of the uninfected cell in respect to the protein specifying mechanism. He proposed that there are actually two components of a double structure complex that specify the amino acid sequence in proteins. Thus in the "recently" divergent proteins, namely: mammalian, most of the replacements in amino acids are not random (X=12.3) due to a result of mutation in one of the components. On the other hand, random (X=1.4) distribution of amino acids in the "anciently" divergent proteins such as insulin, haemoglobin, and the cytochrome Cs is a result of mutations in both components.

It should be noted that Ycas' assignment of amino acids to a particular nucleotide is based on the correlation between the sum of the mole fraction of the amino acids within each group to that of the corresponding nucleotide, within the significance limit of the Bravais-Pearson coefficient (c=0.92, N=24). A serious objection is obviously that the amino acids may be so chosen and assigned to give a good correlation. The results of Yamazaki and Kaesberg (1961) on wild cucumber virus, for instance, have contradicted Ycas' proposal of coding ratio of DNA.

Sueoka (1961a, b) proposed a code of quadruplets or sextuplets, on the basis of GC content of DNA and its correlation with various amino acids. Assume AT and TA to be α and GC and CG to be γ, the quadruplets can be classified into three classes: exclusive (α_4 and γ_4) asymmetric ($\alpha_3\gamma$ and $\alpha\gamma_3$), and symmetric

($\alpha_2\gamma_2$). The frequency of each type of quadruplet will vary with the GC content. The same correlation is expected between amino acids in protein corresponding to each quadruplet and the GC content of DNA. The symmetric class will show no correlation when a linear regression is estimated. Based on the expected correlations within the GC range of analysis (25 to 75%) the best fit is given by assignments in Table 6. Sueoka's model supports the universality of the code among bacteria and probably protozoa also. He suggested that the wide variation of DNA base composition is due to the presence of a large amount of nonsense DNA. He concluded that if the code unit is universal and relatively small (e.g. 3, 4, 5, 6, etc.) it would be even numbers of bases, namely 4 or 6, rather than odd numbers. This hypothesis may be subject to experimental test in time.

Woese (1962) constructed his code (see Table 6) by a correlational study on proteins from six different, but closely related, viruses. The code is characterized by being commafree. He noted (1961) that the replacements in the amino acids of these related viral proteins are not random. Woese further suggested that stereochemically related amino acids correspond to closely related nucleotide code triplets. Smith (1962 b) also suggested that the structure of an amino acid should determine the type of replacement possible. Thus the ones with hydrophobic side chains are replaceable only by another of the same configuration. However, the assumption of Smith that all codes should contain U has no theoretic ground and has been refuted by recent biochemical results.

Jukes (1962 a, b) argued that it is not necessary to expect a stereochemical relationship between the coding triplet and the amino acid. Transfer RNA molecules may originally have been a random series containing all possible coding triplets in combination with all possible amino acid recognition sites. Those that did not provide the correct combinations would have been discarded during the process of evolution. Jukes' model of code assignment is ordered and consistent only with Ochoa's earlier assignment. The data of amino acid replacement in beta-lactoglobulin which he based his calculations upon were shown afterwards to be incorrect (Jukes, 1962 b).

Zubay and Quastler (1962) based on 126 transition replacements of amino acids, proposed a nondegenerate triplet code without specifying the order. They suggested that U is likely to be at one end of each code word. This implies that early in the evolutionary process a dinucleotide code was sufficient, the third nucleotide in a triplet code was added when variation later became desirable. However, their assumption that glu and gln as well as asp and asn are coded by the same triplet may be wrong. If nondegeneracy were the case, the code would have to be highly precise and would be inefficient. The implication that a dinucleotide code was sufficient has also been considered by

Sinsheimer (1959), Roberts (1962) and Petruska (1962) as has been discussed previously.

Hendler (1962) cautioned that a good correlation between the replacement data and a specific code does not necessarily imply a universality nor a complete degeneracy of the code under investigation. This is certainly true. Definite correlation always requires an exhaustive set of data.

c. *Experimental approaches*

There are three main schools of experimentation, all of which have been very important to the coding problem. These are:

1. Chemical approach—feeding a protein-synthesizing system, a known form of synthetic polyribonucleotide, thus mimicing the messenger RNA, and analyzing the polypeptide made.

2. Mutational replacement—invoking specific mutation, by altering a specific form of base pairing, and detecting the amino acid replacements in a specific protein.

3. Genteic approach—analyzing mutations induced at the nucleotide level.

1. The most significant breakthrough in the chemical approach to coding is the studies of Nirenberg and Ochoa and their associates. Ochoa and his colleagues characterized an enzyme, ribonucleotide polymerase, which catalyzes the formation of specific polyribonucleotides. Nirenberg *et al.* (see Martin *et al.*, 1962; Matthaei *et al.*, 1962; Nirenberg *et al.*, 1962) then synthesized a polyuridylic acid and found that it directed the synthesis of polyphenylalanine in a protein synthesizing system of *E. coli*. Such a finding was almost simultaneously made also by Ochoa *et al.* (1961–1962). Both groups conclude that synthetic homo or heteropolynucleotides are active as messenger in stimulating the incorporation of specific amino acids into polypeptides. This is done by introducing into a protein synthesizing system a mixture of all 20 amino acids with one of them labeled with C^{14}. The system is treated with DNAase to inactivate the DNA, thus inhibiting the formation of new messenger RNA. Various synthetic polynucleotides are introduced in place of the messenger RNA. After a period of incubation, trichloroacetic acid- (or tungstic acid) precipitable fractions are collected and the radioactivity measured. In the presence of a polyuridylic acid (poly U; UUUUUUUUUUU....), for instance, most of the activity can be accounted for only due to the incorporation of C^{14} phenylalanine into a polypeptide in the form of Phe-phe-phe-.... They found that there is a positive correlation between the length of the polynucleotide molecule, molecular weight, and the amount of incorporation. Heteropolynucleotides made from various combinations of random mixtures were used and different amino acids were incorporated. Addition of a third base may cause the incorporation of additional amino acids. Recently, Jones and Martin (1962) and

Bretscher and Grunberg-Manago (1962) found that poly CA would also stimulate incorporation of pro and some thr and his. Ochoa *et al.* (1962, VIII) also showed that the composition of the functional synthetic polynucleotides does not necessarily contain U as Zubay and Quastler (1962) and others previously speculated. The codes proposed by the schools of Ochoa and Nirenberg are in part degenerate, in part nonsensical, and has a coding ratio of at least three. The assignment is based on comparison of ratio of the incorporation of phenylalanine to other amino acids in relation to the theoretical proportion of various forms of polynucleotide triplets possible. For instance, if the proportion of U to C to G in a heteropoly UCG is 6:1:1, the proportion of UUU to UCG is expected to be 36, assuming the distribution of various nucleotides in a polymer is random. Now among other incorporations, the activity of phe to ala is 31. In this particular case, 31 is the closest value to 36 and ala is assumed to be incorporated due to UCG (sequence not specified). Such an approximation is very close to the theoretical ones in many cases. They noted that the incorporation of 1 mμ mole of phe into protein requires 1 mμ mole of U in a poly U, suggesting that the function of these polynucleotides is stoichiometric rather than catalytic. The dichotomy of U being excessive in the code but very little in TMV is due to the selective techniques used (see Ochoa *et al.*, VIII). In their work, as discussed above, all assignment is based upon the function of UUU or the incorporation of phe. We know that a linkage of only 4 to 5 molecules of phe is sufficient to form a TCA precipitable fraction. Other amino acid polymers require a longer chain to stay stable in the form of a precipitable polypeptide, and smaller polymers are probably soluble and lost during the washing procedure. Experiments of this type are highly exciting, although there are other pitfalls.

a. The polypeptides so synthesized are not functionally active, thus probably do not exist in nature.

b. The assumption that the distribution of bases in a heteropolynucleotide is at random may be erroneous. DNA tends to consist of blocks of pyrimidines followed by blocks of purines. The deviation of their assignment from theoretical prediction should also indicate this discrepancy.

c. Thus far, only 58 triplets out of a possible 64 were tested and only 41 of them could be assigned with coded information, and only 23 code words are agreeable among these two schools. The sensitivity of detection for certain polypeptides formed has not been totally satisfactory. The conclusions drawn from this technique are thus incomplete. An up-to-date assignment of these groups is included in Table 6.

Even with these apparent shortcomings, the studies of Nirenberg and Ochoa *et al.*, should be considered as the most elegant indirect approach to the coding problem. Some day the correct combination of code may be found by using a

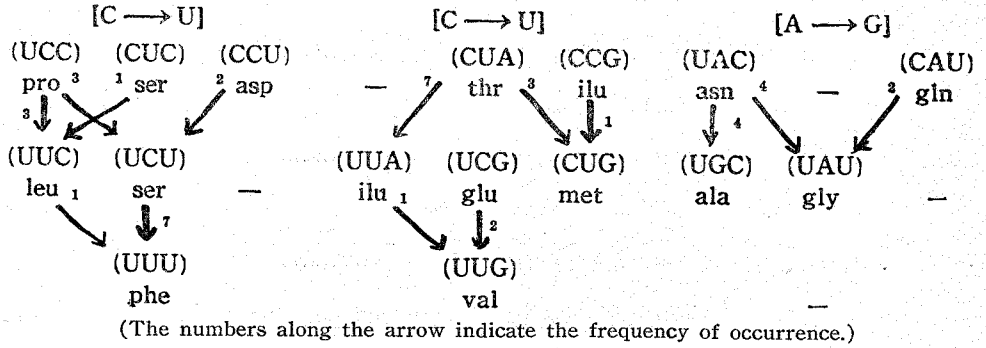
similar technique to this to construct a polynucleotide that would direct the synthesis of a specific protein (e.g., collagen, hemoglobin).

2. Another important experimental approach to the coding problem is that of Tsugita and Fraenkel-Conrat (Tsugita, 1962) and Wittman (1961). Their study is based on the knowledge that the mutagenic action of nitrous acid is through breaking the cross linkage of the helical nucleotides, deaminating the nucleotide bases, and subsequent alteration in base pairing. Thus,

(A) adenine ⟶ hypoxanthine (Hx) ⟶ guanine (G)

(C) cytosine ⟶ uracil (U)

It is assumed that most of the mutations occur by a one hit mechanism, each hit causing a change of only one amino acid in the protein involved. In studying such changes from various nitrous acid induced TMV mutation, Wittman observed that in a number of mutational transitions, the amino acid composition of the mutant protein could be assigned to changes in the bases of the DNA nucleotides. Such an assignment fits well with the octad arrangement of the 64 triplets proposed earlier by Gierer (1961). Thus:



(The numbers along the arrow indicate the frequency of occurrence.)

Only a part of the 6500 nucleotides, probably 500 of one end, is actually involved in the coding of the specific coat protein of TMV investigated. This protein consists of approximately 160 amino acids. The results agree perfectly with the triplet code hypothesis. In fact, as the code is degenerate and non-overlapping, it is compatible with the suggestion of Ochoa *et al.* and Nirenberg *et al.* (loc. cit.). There is also an indication in Wittman's experiment that the decoding process starts from a fixed point.

The study of the A protein of tryptophan synthetase of *E. coli* (Yanofsky *et al.*, 1961; Helinski and Yanofsky, 1962; Henning and Yanofsky, 1962) is also very interesting. Genetic studies showed that two UV induced mutants, each altered in one amino acid, are very closely linked. One has its gly replaced by arg and the other by glu. Applying the code agreed upon by Ochoa and Nirenberg's groups, the replacements can be shown as:

...GUG...⟶...GUC...
(gly) (arg)
...GUG...⟶...GAG...
(glu)

thus a recombination between GUC and GAG at the position between UC or AG would yield a wild type GUG, with GAC (either ser or asp) as the simultaneous product. This result may well serve as a good illustration of recombination at the molecular level. When an exhaustive set of such controlled mutations is available, we may be able to see more clearly the coding system involved. At least it should show whether a congruency exists between the nucleotide bases (gene) and the amino acid sequence in that protein produced. A more complete replacement data including N-bromosuccinimide and dimethylsulphate induced TMV mutants of Tsugita and hemoglobin mutations of Smith and Perutz have been summarized by Crick (1962).

3. The genetic experiments of Crick *et al.* (1961) are highly significant to the coding problem. It is known that proflavin acridine causes mutation either by addition or deletion of a single nucleotide base in DNA through inserting itself between two adjacent nucleotide pairs and increasing their distance to

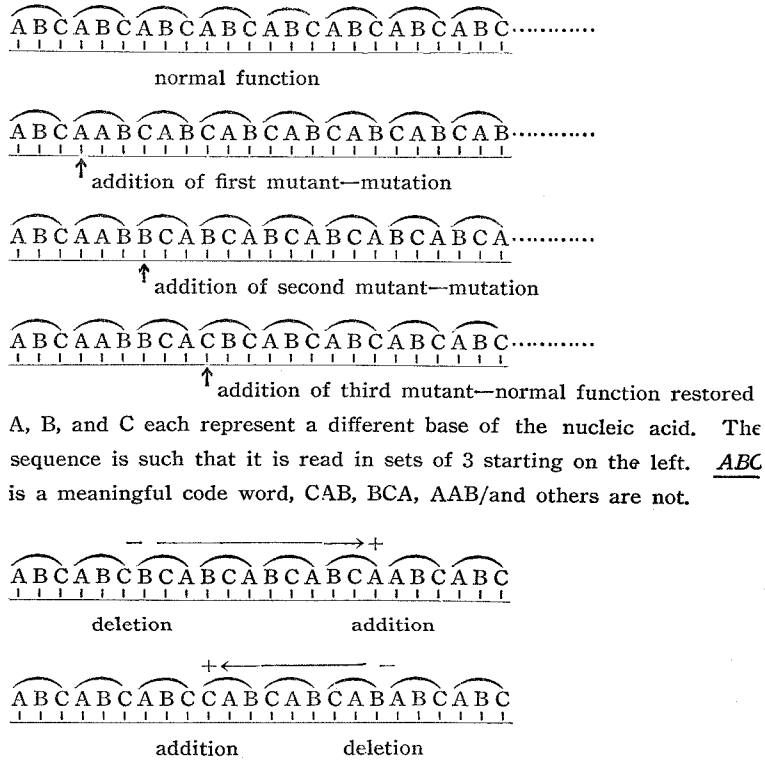


Figure 1

A, B, and C each represent a different base of the nucleic acid. The sequence is such that it is read in sets of 3 starting on the left. *ABC* is a meaningful code word, CAB, BCA, AAB/and others are not.

6.8 Å instead of 3.4 Å. In coliphage a marker gene rII is known to have a definite phenotype. Thus by inducing mutations in coliphage specifically in the rII region, Crick *et al.* were able to construct triple mutants having 2 forms, involving either 3 additions or 3 deletions. Taking a triple set of mutants, either of the addition or deletion type, they found that either mutating singly or in pairs resulted in an absence of function in the gene. However, when the three mutants were combined in the same genome, the function of rII was restored or partly restored. Combination of a single deletion and a single addition also restores the function (see Figure 1). Thus they suggest that the coding ratio is 3 or a module of 3. They also suggest that decoding synchronization is achieved by starting at one end of the message and reading off 3 bases at a time without overlapping. They infer that the code is degenerate or generally more than one triplet code for each amino acid.

Generalities of a Plausible Code

In the review made above, the following generalities of a code have been stressed: the length of a code and its uniformity, the arrangement of the codes, overlapping versus nonoverlapping; the nature of the codes, degenerate, ambiguous or nonsensical; synchronization of the decoding process; and the universality of the code. One may now proceed to discuss the findings collectively under these headings.

A. *Code length*

As there are 4 types of bases and 20 or so amino acids, only 4 major models of correspondence have been constructed, namely:

a. Each base codes for a certain group of amino acids——coding ratio of 1 (Ycas).

b. A combination of two bases corresponds to a certain amino acid——duplet code (Sinsheimer, Roberts).

c. Three bases, same or different, arranged in a linear array and coding as an entity, for one amino acid——triplet code (Crick, Wittman, Woese, Ochoa, Nirenberg).

d. More than 3 bases are involved in each coding unit, however, operating in a similar fashion as triplets (Sueoka, Golomb).

While a coding ratio of 3 is the most favored model today, the actual unit length of a code is yet to be determined. Rough estimates of the amount of DNA necessary to code for a single amino acid have been done by Levinthal (1959) and Garen (1960). The result is consistent with a coding ratio of 3 to 10 as obtained independendently by Fuerst, Wollman and Jacob (1959) based on recombination length of DNA and genetic and chemical studies with alkaline

phosphatase. Ofengand and Haselkorn (1961) also conclude that the coding ratio is lower than 9 but higher than 3 from their work with viral RNA-dependent incorporation of amino acids.

Also, most of the coding models agree upon a uniform length. But, this is based purely on mathematical convenience. Blumenthal (1962), for instance, reasoned that a triplet code does not contain enough information. Levinthal (1959) suggested that other models of code may exist. For example it may be that every 10th nucleotide (one complete turn of a helical DNA molecule) specifies the specific amino acid. Roberts (1962a) did indicate in passing that codes of unequal length may exist and function simultaneously. A complex coding system of this type would require a much more precise and complicated system of programming. Nevertheless, there is no evidence so far against this possibility. Crick (1962) considered it a possibity to have triplets code for the common amino acids and doublets for the rarer ones. He even suggested that the length of a code does not need to be an interger.

B. *Nonoverlapping vs. overlapping*

Wall (1962) advocates that the genetic code is overlapping instead of non-overlapping as popularly believed. Indeed, the results of Crick *et al.* (1961), for instance, can be explained equally well by this overlapping scheme. This scheme is characterized by a code length not exceeding 5 and a coding ratio of either 3 or 4. Wall reasoned that there are probably two types of mutation, missense and nonsense. Nonsense mutation results in either lethal or undetectable changes. Missense mutation leads to replacements of amino acids in certain proteins. The site of a mutation may or may not lie in the overlapping region. Therefore, this model is compatible with Crick's supposition of deletion or addition of nucleotides in proflavin-induced mutants. Wall pointed out that certain combinations of amino acid sequence (altogether there are 20^2 possible nearest neighbor combinations; 20^3 for a series of three..) are not found because of a non-ergodic characteristic of the amino acid sequential process. Wall's non-ergodic hypothesis seems to be a special case of the work of Rosen (loc. cit.) as previously discussed. Rosen suggested that the DNA-protein coding processes in nature are actually quasi-ergodic, under which a finite number of DNA-protein codes are possible. Biochemical experiments alone are insufficient to resolve this aspect of problem, although Goldstein (1962) based on results from protein synthesis suggested that a stepwise assemblage process could preclude errors in decoding and allow the existence of an overlapping code.

C. *Degeneracy, ambiguity and nonsense*

If the code is indeed triplet in nature, it is easier to visualize that some codes are degenerate, others ambiguous, and still others nonsensical. Ambiguity of a code means one code word for more than 1 amino acid. Matthaei *et al.*

(1962) and Bretscher and Grunberg-Manago (1962) both observed that poly U could code for leucine as well as phenylalanine *in vitro*. Pleitropism of a gene, for example, could be an analogy at a higher level. The fact that a certain genetic region is profound in function or is a mutational cold spot may be an indication that nonsense exists. There is also the recent direct evidence of Crick *et al.* (1961) to support this.

The hypothesis that degeneracy, at least partially, does exist has received support from the mutational replacement results of Wittman (1961), correlation studies of Reichmann *et al.* (1962), and the chemical experiments of Nirenberg *et al.* and Ochoa *et al.* Reichmann *et al.* (1962) found that the distribution of C/U in the end fractions (13.2% molar of the total digest) in the tobacco necrosis virus is inadequate to account for the information required to code for the proteinous satellite components. They assumed that for 240 amino acid residues, the estimated protein composition of the satellite, 720 nucleotides or 60% of the total RNA would be required on a triplet code basis and that there are about 387 triplets for the total length of this viral RNA. Percent distribution of U:C:A:G for this RNA is respectively 24.9:22.1:28.0:25.0. Weisblum *et al.* (1962) too have provided physical bases for such a property, at least with leucine. The incorporation of leucine can be stimulated equally well by poly UC and poly UG. Moreover, it is likely that the priming ability of different code words is different. It would be interesting to know, for example, what is the maximal degree of degeneracy possible, whether chemically similar amino acids have the same degree of degeneracy, and whether certain types of triplets are more vulnerable to ambiguity or nonsense than others. That at least 23 code assignments have been agreed upon from biochemical evidence leaves little doubt that some codes are highly degenerate. It is interesting to note that the degeneracy is not at random (Table 6).

D. *Orientation of the decoding process*

So far the only direct experimental suggestion for an orientation during the decoding process comes from studies of Crick *et al.* (1961), Wittman (1961). They suggest that the sequence of code is deciphered in one direction along the strand starting at a fixed point. There is biochemical evidence for the oriented replication of DNA on a pre-existing DNA (Yoshikawa and Sueoka, 1962), but it is unknown how the messenger RNA is assembled on the priming DNA. Dintzis (1961), Goldstein (1962) and others have shown that some proteins are assembled starting at the N-terminal end; the decoding process from messenger RNA may also be oriented and synchronized. They have evidence that there is only one correct polarity for the peptide-bond formation. Recognition may be achieved by having an alpha-amino acid to match the possibly phosphorylated 3′ and 5′ hydroxyl group at the terminal end of the messenger RNA. A variety of RNA

specifying various amino acids may be involved in the recognition process. However, once a unique terminal group was established on the template, the condition would be sufficient to initiate and to program the growth of a protein chain. This idea is parallel to the consideration of Pattee (loc. cit.) as discussed earlier.

More recent experiments of Ochoa's group (1962, VII) using synthetic polymer AUUUUU....U, noted that tyr (coded by UUA) was present not at the N terminal (3′ hydroxyl), but the C terminal (carboxyl) end of the polypeptide formed. Similar results were obtained with GUUUUU....U, providing a very strong supportive evidence for a directional decoding process.

The suggestion that a full stop at the end of a coding sequence for a cistron may have been provided by the discovery of rare and novel bases in the DNA nucleotide. Dunn and Smith (1958) for example have found that there exists one residue of 6-methylaminopurine in every 250 nucleotides in *E. coli,* and one in every 800 in T2 coliphage DNA. This 6-methylaminopurine may be nonsensical and either stop or break a sequential reading process. Crick (1962) hinted that two letters in a code word need not be situated adjacently.

E. *Universality*

In all the organisms studied, evidence has been found for a stepwise control mechanism for the synthesis of DNA-RNA-protein. However, uncertainty remains as to the universality of a general coding system. Crick (1962), Champe and Benzer (1962), and Yanofsky *et al.* (1961) believe that there are discrete differences in the method of coding in different organisms due to variations in the specificity of the activating enzymes. Ochoa *et al.* (1962) believe that the code is universal. Brenner (1961) also suggests that the coding mechanism is unitary in nature. Rolfe and Meselson (1959) suggested that individual amino acid coding sequence may be species-specific. The genetic factors controlling the synthesis of β-galactosidase and alkaline phosphatase have also been found to be active in both *E. coli* and *Serratia* interchangeably (Singer *et al.,* 1961). Ehrenstein and Lipmann (1961) reported the synthesis of hemoglobin very similar to that of rabbits using rabbit ribosome and mRNA and tRNA from either *E. coli,* yeast, or *Micrococcus lysodeikticus.* Similar observations have also been reported by Niu *et al.* (1962). Controlled synthesis of polypeptides has also been demonstrated by Arnstein *et al.* (1962) and Bretscher and Grunberg-Manago (1962) in cell-free systems other than the ones used by the schools of Nirenberg and Ochoa. Tsugita, Fraenkel-Conrat, Nirenberg and Matthaei (1962) have shown recently that a protein similar to tobacco mosaic viral protein can be synthesized in a cell-free protein synthesizing system of *E. coli* when the infective tobacco mosaic virus RNA is introduced. It may be safe to conclude that certain codes may be equally functional in different organisms.

Summary

To recapitulate, the basic concepts involved today in encoding and decoding genetic information are as follows:

a. DNA is the primary vehicle of inheritance and contains the coding information. Some of the genetic information, if not all, is to specify the sequence of amino acids in proteins.

b. Messenger RNA, synthesized on, and having a homologous base sequence to DNA, transcribes the genetic information.

c. The decoding of genetic information involves the assemblage of a specific sequence of amino acids into a protein molecule mediated by a protein synthesizing system.

d. The sequence of nucleotide bases on the messenger RNA reflects that of DNA and corresponds to the amino acid sequence of the protein produced.

e. Coding is an inferential design for the understanding of basic correlations between the chemical structures of genetic material and its products.

Although there are a few generalities that can be drawn from the recent experimental approaches, the problem of coding is still open to a series of questions:

- Is the uniform code length valid?
- Is degeneracy of the codes an exception or the rule?
- Is the degeneracy distributed at random among various code words?
- Is the messenger read sequentially from a fixed point?
- Is the message polarized?
- Is the code indeed nonoverlapping and commafree?
- Is there one and only one coding system in each organism?
- Is the coding system universal?
- Is the decoding process in nature fallible? If so, what type of mechanism is provided for correction?
- Is it possible to have a sequential code built upon information feedback?
- Is there a secondary control mechanism, supplementing or interacting with the genetic code?

In time, some of these questions will be answered while others may remain obscure. It would be informative to know the activity of other synthetic polymers such as GUUUUU....U and AUUUUU....U as tested by Ochoa *et al.* recently. Copolymers with one or two or three base nucleotides at one end differing from the rest of the polymer should further reveal the property of a code regarding its orientation, efficient length, overlapping, synchronization, and possibly degeneracy. This may require a small size of polyribonucleotide with the total number of bases being a multiple of three. On the other hand, fractionation of functional DNA could also be very useful in understanding the

coding nature. Progress has been made, for example, with the isolation of the pyrimidine segment of a single stranded DNA from a coliphage ϕX 174. The longest fraction so far isolated contained 11 pyrimidine base consecutively. Also two chains of 10 and four of 9 have also been fractionated (Hall, 1962). The determination of the exact composition as well as the biological function of these chains is still in progress. Jones and Nirenberg (1962) reported, however, that longer chains (>100 units) of synthetic polymers are more active than the shorter ones.

Similarly, the use of polynucleotides substituted with base analogs in cell free studies may also augment our uderstanding of the codes. Poly UG after nitrous acid treatment would be changed to UX (xanthin) and is greatly inactivated (Jones and Nirenberg, 1962). Basilio *et. al.* (see Ochoa *et al.*, 1962, VIII) showed that UH (hypoxanthin) acts similarly to UG *in vitro*. These results are compatible with DNA synthetase and RNA polymerase behaviors. Also poly UI (inosine) is as effective as poly UG, but poly UX is not; poly CU is similar to CG in incorporation activity. It is unclear as to why 5FU (fluorouracil) and N-methyl uracil (MeU) homopolymers are inactive. 5FU when copolymered with U is active, but not MeU with U. Haschemeyer and Rich (1962) also reported that poly T (thymidylic acid) is inactive for the incorporation of phe, probably due to the competitive occupancy of the active sites. These studies are essential to the understanding of mutagenic as well as the encoding mechanisms. They have suggested at least that under certain conditions, some codes could easily become nonsense, if they were not nonsense, if they were not nonsensical to begin with.

Sueoka and Cheng (1962) have isolated a naturally existing deoxypoly A-T in a marine crab (*Cancer borealis*), indicating that heteropolynucleotides of this type may in fact exist in nature and function as the synthetic ones.

From another point of view, genetic regions known to control the synthesis of a specific protein may eventually be isolated, and the nucleotide base sequence of the genetic region and the amino acid sequence of the protein involved, compared. The isolation of a particular genetic segment, the rII region of coliphage T4, of known function has been accomplished by Hall and Spiegelman (1962), Bautz and Hall (1962).

Newer techniques have also been developed recently to determine the base sequences of various nucleotides. Beer and Mondrianakia (1962) took the advantage of greater electron density of 8-amino-1, 3, 6-naphthalene-trisulfonic acid to locate various bases in a DNA molecule using the electron microscope. Wilska (1962) is constructing an electron microscope having a resolving power of 2 to 3 Å. Champe and Benzer (1962) studied the base sequence of messenger RNA with 5-fluorouracil (5FU). 5-FU occasionally pairs as C and will substitute

for U in a messenger RNA. Thus if a mutation exists such that a C-G pair in DNA is changed to T-A, the corresponding messenger RNA would be changed from C to U and would pair in the complementary tRNA with A instead of G. When 5-FU replaces U the phenotype will be restored. Knowing the mutational sites involved by recombination test, Champe and Benzer (loc. cit.) have been able to construct a map of messenger RNA for the rII region of the coliphage T4 in terms of nucleotides for the standard type. As 5-FU involves the detection of only G and C, the positions of the other bases will be determined upon the development of analogues of C or G that function similarly to 5-FU in replacing U. Unfortunately, the specific enzyme or enzymes involved in the function of rII are still unknown. It probably involves a protein not contained in the phage itself, but is required in the infective cell for the formation of new phage. In time, it is hoped that some ingenious techiques will be developed in systems with defined gene products such as tyrosinase (Horowitz *et al.*, 1961) and the hypothesis of correspondence between the base sequence and amino acid sequence examined.

Clearly, the subject of coding is being advanced very rapidly especially in the past year. More experimental results, both genetical and biochemical, are to come. It is not only an opportune time to reappraise and re-evaluate critically the theoretical foundations as well as the experimental results, but also a high time to be optimistic about the understanding that we shall be led to. It is hoped that this paper serves to provide the view of a general, and most likely correct, trend of approaches to the understanding of the problem on encoding and decoding of genetic information. A true and plausible genetic code may well be a modified form limited by the generalities outlined above.

Before concluding, it should be pointed out that the present thesis of coding is based on DNA being the sole vehicle of inheritance. Such a basis is subject to a few but obstinate and unreconcilable exceptions. Lindegren (1961) suggested that the role of DNA is merely for the assurance of synapsis during meiosis and of proper configuration of the genetic material. This is based on his observation that the most of DNA is not located on "genes." The acceptance of his suggestion certainly depends upon the concept of "genes" —a term which is being constantly reappraised.

Also, the cytoplasm of a cell has been shown to be decisive and intriguing in determining gene functions, even though the nature of its role is still unclear. Furthermore, there may exist chromosomal and cytoplasmic elements like episomes (Jacob and Wollman, 1961) which may possess a maneuverable rather than steady coding system. Still others, McCully and Cantoni (1962) lately hypothesized that RNA is synthesized by a template mechanism independent of DNA, calling for the broadening of our concept that genetic information is

transmitted only from DNA (see also Chapevielle *et al.*, 1962). Sager (1959) postulated that there is an autonomous genetic RNA and that DNA contributes only a small although critical part of the total information. Experiments of Shen *et al.* (1961) on RNA being a transforming principle in *Bacillus subtilis* seems to provide a supporting evidence to her hypothesis. Nonetheless, we believe that the correct trend of approaches to the problem have been outlined in this paper and that all functions of an organism may be expected to converge on a master control in DNA by way of the process of encoding and decoding.

Acknowledgments

The authors wish to express their sincerest appreciation for the fine academic atmosphere and stimulating discussions maintained by their colleagues at the California Institute of Technology. They wish to acknowledge Professors Sterling Emerson, James Bonner, Norman Horowitz, and Mr. John Urey for their encouragement and valuable suggestions during the preparation of this manuscript; to Professor Robert Sinsheimer for providing several preprints that he received. This manuscript is prepared with supports from the U.S. Public Health Service, National Institutes of Health, Grants No. GM06965 and RG5143.

遺傳情報之譯傳

黃秉乾　黃周汝吉

遺傳因子之組成物質爲去氧核糖核酸*，其證據很多。近年研究結果，公認遺傳情報之譯載及傳遞，是經由核糖核酸。至蛋白質之結構成分及性能表達，乃係由於蛋白質之形成受核糖核酸之支配，而某一種核糖核酸（所謂傳達核糖核酸）之成分與去氧核糖核酸又相駢對之故。前人復指出細胞代謝、分化、成熟及衰老等過程，皆係由蛋白質酵素爲媒介。因此，追究核酸與蛋白質結構上之相關，乃成爲追究遺傳因子與個體性狀之關鍵。此一問題之探討，已成爲較近分子生物學研究最熱烈的目標之一。

目下有關遺傳情報譯傳之理論，似以"三分子"學說最爲成熟。此學說是假定蛋白質與核酸成分間有正相關關係。蛋白質之性能係由其分子構造決定，而其分子構造則由所含廿種

*註：

去氧核糖核酸	deoxyribonucleic acid	氨基酸	amino acid
核糖核酸	ribonucleic acid	脯氨酸	proline
核氮基	nucleotide base	蛋氨酸	tyrosine
C核氮基	cytosine	天門冬酸	aspartic acid
A核氮基	adenine	苯丙氨酸	phenylalanine
U核氮基	uracil	三分子學說	triplet theory
G核氮基	guanine	微粒	virus

氨基酸之比例及排列順序而異。核酸之結構較爲規律，惟其所含四型核氮基之序次亦多變化。三分子學說假定：每三分子的核氮基（不問其爲同型或異型）是決定着一種氨基酸，而廿種氨基酸是由六十四種三分子排列中之若干不同組合所決定。近年來蛋白質合成之研究大有進展，並已知在試驗管中合成蛋白質，不僅藉酵素之作用，且亦受傳息核糖核酸之支配。若該合成核酸僅含一型核氮基，其所促成之蛋白質性物質亦僅含有一種氨基酸。增加一型核氮基時，則多一種或數種氨基酸。例如某一種核糖核酸純由C核氮基酸組成，其所產生者乃爲一多分子腩氨酸。若加入少量A核氮基，則可發現合成品中雜有蛋氨酸。又如U氮基酸爲該核酸之唯一成分時，所產生者僅有苯丙氨酸。再加入G核氮基及U核氮基時，則合成品成分中含有天門冬酸及其他數種氨基酸。此皆視三型核氮基之使用比例如何而定。據推算，大致每一分子氨基酸之組成，是需要三個分子的核氮基。說明核酸之成分可直接決定蛋白質之成分，是三分子學說之極有力的生物化學的證據。

三分子及其他譯傳學說之由來不久，各學派之意見亦多揣測與分岐，惟譯傳之基本理論，則除上述例證外，已獲微粒遺傳、化學誘變、蛋白質及核酸之定量分解、生物物理學、分子遺傳學以及統計數理上之支援。

本文是就近年來該一學說的發展及作者在該一方面的心得而作成的綜合性論述。(摘要)

Literature Cited

AGENO, M. Deoxyribonucleic acid code. Nature **195**: 998-999, 1962.

ARNSTEIN, H. R. V., R. A. COX, and J. A. HUNT. Function of polyuridylic acid and ribonucleic acid in protein biosynthesis by ribosomes from mammalian reticulocytes. Nature **194**: 1042-1044, 1962.

BAUTZ, E. K. F., and B. D. HALL. The isolation of T4 specific RNA on a DNA-cellulose column. Proc. Nat. Acad. Sci. U. S. **48**: 400-408, 1962.

BEADLE, G. W., and E. L. TATUM. Genetic control of biochemical reactions in *Neurospora* Proc. Nat. Acad. Sci. U. S. **27**: 499-506, 1941.

BEER, M., and E. N. MOUDRIANAKIS. Determination of base sequence in nucleic acids with the electron microscope: visibility of a marker. Proc. Nat. Acad. Sci. U. S. **48**: 409-416, 1962.

BENZER, S. The elementary units of heredity. In Chemical Basis of Heredity (McElroy & Glass, eds.) pp. 70-93, 1957.

BONNER, J., R. C. HUANG, and N. MAHESHWARI. The physical state of newly synthesized RNA. Proc. Nat. Acad. Sci. U. S. **47**: 1548-1554, 1961.

BLUMENTHAL, L. K. A contribution to the coding problem. J. Theoret. Biol. **2**: 72-73, 1962.

BRENNER, S. The impossibility of an overlapping triplet code in information transfer from nucleic acid to proteins. Proc. Nat. Acad. Sci. U. S. **43**: 687-694, 1957.

BRENNER, S. RNA, ribosomes, and protein synthesis. Cold Spr. Harb. Symp. Quant. Biol. **26**: 101-110, 1961.

BRENNER, S., F. JACOB, and M. MESELSON. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature **190**: 576-581, 1961.

BRETSCHER, M. S., and M. GRUNBERG-MANAGO. Polyribonucleotide-directed protein synthesis using an *E. coli* cell-free system. Nature **195**: 283-285, 1962.

CHAMBERLIN, M., and P. BERG. DNA-directed synthesis of RNA by an enzyme from *E. coli*. Proc. Nat. Acad. Sci. U. S. **48**: 81-94, 1962.

CHAMPE, S. P., and S. BENZER. Reversal of mutant phenotype by 5-fluorouracil: an approach to nucleotide sequences in messenger RNA. Proc. Nat. Acad. Sci. U. S. **48**: 532-546, 1962.

CHAPEVIELLE, F. *et al.* On the role of soluble RNA in coding for amino acids. Proc. Nat. Acad. Sci. U. S. **48**: 1086-1092, 1962.

CHAVCHARIDZE, V. V. The primary "alphabet" of deoxyribonucleic acid. Biofizika **3**: 391-395 (Biophysics **3**: 377-381), 1958.

CRICK, F. H. C. The recent excitement in the coding problem to appear in Progress in Nucleic Acid Research (Davidson & Cohn, eds.), 1962. [in press]

CRICK, F. H. C., J. S. GRIFFITH, and L. E. ORGEL. Codes without comma. Proc. Nat. Acad. Sci. U. S. **43**: 416-421, 1957.

CRICK, F. H. C., L. BARNETT, S. BRENNER, and J. WATTS-TOBIN. The general nature of the genetic code. Nature **192**: 1227-1232, 1961.

DINTZIS, H. M. Assembly of the polypeptide chains of hemoglobin. Proc. Nat. Acad. Sci. U. S. **47**: 247-261, 1961.

DOUNCE, A. L. Duplicating mechanism for peptide chain and nucleic acid synthesis. Enzymologia **15**: 251-258, 1952.

DUNN, D. B., and J. D. SMITH. The occurrence of 6-methylaminopurine in deoxyribonucleic acid. Biochem. J. **68**: 627-636, 1958.

ECK, RICHARD V. The protein cryptogram: I. Non-random occurrence of amino acid "alleles." J. Theoret. Biol. **2**: 139-151, 1962.

EHRENSTEIN, G. VON, and F. LIPMANN. Experiments on hemoglobin biosynthesis. Proc. Nat. Acad. Sci. U. S. **47**: 941-950, 1961.

FRENDENTHAL, HANS. Ein kombinatorisches Problem von biochemischer Herkunft. K. Ned. Akad. Wetenschap. Proc., Series A, Mathematical Sciences **61**: 253-258, 1958.

FREESE, E. On the evolution of the base composition of DNA. J. Theoret. Biol. **3**: 82-101, 1962.

GAMOV, G., A. RICH, and M. YCAS. The problem of information transfer from the nucleic acids to proteins. Adv. Biol. Med. Phys. **4**: 23-68, 1956.

GAREN, A. Genetic control of the specificity of the bacterial enzyme, alkaline phosphatase. In Microbial Genetics (W. Hayes & R. C. Clomes, eds.) pp. 239-247, 1960.

GIERER, A. Proc. Vth Int. Congr. Biochem. Moscow (1961) Symp. III (From Wittman 1961), 1961.

GOLDSTEIN, A. Chain growth of proteins—some consequences for the coding problem. J. Molec. Biol. **4**: 121-122, 1962.

GOLOMB, S. L., R. WELCH, and M. DELBRÜCK. Construction and properties of commafree codes. K. Danske Vidensk, Selsk (Biol. Medd.) **23**: 9, 1958.

GOLOMB, S. Personal communication, 1962.

HALL, J. B. Pyrimidine oligonucleotides of the DNA of the bacteriophage ϕX 174. Biology Annual Report, C. I. T., pp. 26-27, 1962.

HALL, B. D., and L. S. SPIEGELMAN. Sequence complementarity of T2-DNA and T2 specific RNA. Proc. Nat. Acad. Sci. U. S. **47**: 137-146, 1962.

HARTMAN, P. E. Transduction: A comparative review in Chemical Basis of Heredity (McElroy & Glass, eds.), Johns Hopkins Press, 1957.

HASCHEMEYER, A. E. V., and A. RICH. Investigations on the polyuridylic acid-dependent stimulation of phenylalanine incorporation in *Escherichia coli* cell-free systems. Biochim. Biophys. Acta **55**: 994-997, 1962.

HELINSKI, D. R., and C. YANOFSKY. Correspondence between genetic data and the position of amino acid alteration in A protein. Proc. Nat. Acad. Sci. U. S. **48**: 173-183, 1962.

HENDLER, R. W. On the agreement of amino acid replacement data with code designations for the amino acids. Proc. Nat. Acad. Sci. U. S. **48**: 1402-1408, 1962.

HENNING, U., and C. YANOFSKY. An alteration in the primary structure of A protein predicted on the basis of genetic recombination data. Proc. Nat. Acad. Sci. U. S. **48**: 183-190, 1962.

HERSCH, R. T. Mutants of TMV and the commaless code. J. Theoret. Biol. **2**: 326-328, 1962.

HOROWITZ, N. H., M. FLING, H. L. MACLEOD, and Y. WATANABE. Structural and regulatory genes controlling tyrosinase synthesis in *Neurospora* Cold. Spr. Harb. Symp. Quant. Biol. **26**: 233-238, 1961.

HOTCHKISS, R. D. Criteria for quantitative genetic transformation of bacteria in Che. Basis of Heredity (McElroy & Glass, eds.), Johns Hopkins Press, pp. 34-335, 1957.

HUANG, R. C., and J. BONNER. Histone, a suppressor for RNA synthesis. Proc. Nat. Acad. Sci. U. S. **48**: 1216-1222, 1962.

HURWITZ, J. *et al.* The enzymatic incorporation of ribonucleotides into RNA and the role of DNA. Cold. Spr. Harb. Symp. Quant. Biol. **26**: 91-100, 1962.

INGRAM, V. M. Gene mutation in human haemoglobin: the chemical difference between normal and sickle-cell haemoglobin. Nature **180**: 326-328, 1957.

ITO, E., and J. L. STROMINGER. Enzymatic synthesis of the peptide in a uridine nucleotide from *Staphylococcus aureus*. J. Biol. Chem. 235, PC 5-7, 1960.

JACOB, F., and J. MONOD. Genetic regulatory mechanism in the synthesis of proteins. J. Molec. Biol. **3**: 318-356, 1961.

JACOB, F., and E. L. WOLLMAN. Les episomes, elements genetiques ajoutes. C. R. Acad. Sci., Paris, **247**: 154-156, 1958.

JONES, D. W., and A. G. MARTIN. Composition of genetic coding units. Fed. Proc. **21**: 414, 1962.

JONES, D. W., and M. N. Nirenberg. Qualitative survey of RNA code words. Proc. Nat. Acad. Sci. U. S. **48**: 2115-2123, 1962.

JOSSE, J., D. KAISER, and A. KORNBERG. Enzymatic synthesis of DNA. VIII. Frequency of nearest neighbor base sequences in DNA. J. Biol. Chem. **236**: 864-875, 1961.

JUKE, THOMAS H. Beta lactoglobulins and the amino acid code. Biochem. Biophys. Res. Comm. **7**: 281-283, 1962a.

JUKE, THOMAS H. Possible base sequence in the amino acid code. Biochem. Biophys. Res. Comm. **7**: 497-502, 1962b.

LEDLEY, R. S. Digital computational methods in symbolic logic with examples in biochemistry. Proc. Nat. Acad. Sci. U. S. **41**: 498-511, 1955.

LEVINTHAL, C. Coding aspects of protein synthesis. Rev. Modern Physics **31**: 249-255, 1959.

LEVINTHAL, C. Genetic and chemical studies with alkaline phosphatase of *E. coli*. Brookhaven Symp. Biol. **12**: 35-39. 1959.

LINDEGREN, C. C. The biological function of deoxyribonucleic acid. J. Theoret. Biol. **2**: 107-119, 1961.

MARTIN, R. G., J. H. MATTHAEI, O. W. JONES, and M. W. NIRENBERG. Ribonucleotide composition of the genetic code. Biochem. Biophys. Res. Comm. **6**: 410-414, 1962.

MATTHAEI, J. H., O. W. JONES, R. G. MARTIN, and M. W. Nirenberg. Characteristics and composition of RNA coding units. Proc. Nat. Acad. Sci. U. S. **48**: 666-677, 1962.

MCCALLY, K. S., and G. L. CANTONI. Non-random base sequence of sRNA and an hypothesis for sRNA biosynthesis. J. Molec. Biol. **5**: 80-89, 1962.

MORGAN, T. H. The theory of the gene. Yale Univ. Press, 1928.

NIRENBERG, M. W., J. H. MATTHAEI, and O. W. JONES. An intermediate in the biosynthesis of polyphenylalanine directed by synthetic template RNA. Proc. Nat. Acad. Sci. U. S. **48**: 104-109, 1962.

NIU, M. C. *et al.* RNA-induced biosynthesis of specific enzymes. Proc. Nat. Acad. Sci. U. S. **48**: 1964-1969, 1962.

OCHOA, S. *et al.* (1961-1962). Synthetic polynucleotide and the amino acid code, Proc. Nat. Acad. Sci. U. S.

I. P. Lengyel, J. G. Speyer, and S. Ochoa **47**: 1936-1942, 1961.
II. J. F. Speyer, P. Lengyel, C. Basilio, and S. Ochoa **48**: 63-68, 1962.
III. P. Lengyel, J. F. Speyer, C. Basilio, and S. Ochoa **48**: 282-284, 1962.
IV. J. F. Speyer, P. Lengyel, C. Basilio, and S. Ochoa **48**: 441-448, 1962.
V. C. Basilio, A. J. Wahba, P. Lengyel, J. F. Speyer, and S. Ochoa **48**: 613-616, 1962.
VI. A. J. Wahba, C. Basilio, J. F. Speyer. P. Lengyel, R. S. Miller, and S. Ochoa **48**: 1683-1686, 1962.

VII. R. S. Gardner, A. J. Wahba, C. Basilio, R. S. Miller, P. Lengyel, and J. F. Speyer **48**: 2087-2094, 1962.

VIII. A. J., Wahba, R. S. Gardner, C. Basilio, R. S. Miller, J. F. Speyer, and P. Lengyel 1963 (in press)

OFENGAND, J., and R. HASELKORN. Viral RNA-dependent incorporation of amino acid into protein by cell-free extracts of *E. coli*. Biochem. Biophys. Res. Comm. **6**: 469-474, 1962.

PATTEE, H. H. On the origin of macromolecular sequences. Biophys. J. **1**: 683-710, 1961.

PETRUSKA, J. Possible nature of the amino acid code. Biology Annual Report, C. I. T., p. 35, 1962.

REICHMANN, M. E. *et al.* Experimental evidence for the degeneracy of the nucleotide triplet code. Nature **195**: 999-1000, 1962.

RILEY, M., and A. B. PARDEE. Gene expression: its specificity and regulation. Ann. Rev. Microbiol. **16**: 1-34, 1962.

ROBERTS, R. B. Alternative codes and templates. Proc. Nat. Acad. Sci. U. S. **48**: 897-900, 1962a.

ROBERTS, R. B. Further implications of the doublet code. Proc. Nat. Acad. Sci. U. S. **48**: 1245-1250, 1962b.

ROLFE, R., and M. MESELSON. The relative homogeneity of microbial DNA. Proc. Nat. Acad. Sci. U. S. **45**: 1039-1042, 1959.

ROSEN, R. A quantum-theoretic approach to genetic problems. Bull. Math. Biophys. **22**: 227-255, 1960.

ROSEN, R. An hypothesis of Freese and the DNA-protein coding problem. Ibid. **23**: 393-404, 1961a.

ROSEN, R. On the role of chemical systems in the microphysical aspects of primary genetic mechanisms. Ibid. **23**: 393-404, 1961b.

SAGER, R. Discussion following Crick's paper on the present position of the coding problem. Brookhaven Symp. in Biol. **12**: 39, 1959.

SHEN, S. C., M. M. HUNG, S. C. TSAI, H. C. CH'EN, and W. Y. CHANG. Ribonukleinovaya kislota kak transformirovannyi zlement v bakterii (Ribonucleic acid as a genetic transforming principle in bacteria). Koo Hsueh T'ung Pao **16**: 491-494, 1960. Also Chemical Abstracts **55** (18): 17747a, 1961.

SINGER, G. R. *et al.* Gene expression in intergenic merozygotes. Cold. Spr. Harb. Symp. Quant. Biol. **26**: 31-34, 1961.

SINSHEIMER, R. L. A single-stranded DNA from bacteriophage ϕX 174 in structure and function of genetic elements. Brookhaven Symp. in Biol. **12**: 27-34, 1959.

SINSHEIMER, R. L. Is the nucleic acid message in a two-symbol code? J. Molec. Biol. **1**: 218-220, 1959.

SMITH, E. L. Nucleotide base coding and amino acid replacements in protein. I. Proc. Nat. Acad. Sci. U. S. **48**: 677-684, 1962a. II. **48**: 859-864, 1962b.

SPIEGELMAN, S. The relation of informational RNA to DNA. Cold. Spr. Harb. Symp. Quant. Biol. **26**: 75-90, 1961.

SUEOKA, N. Variation and heterogeneity of base composition of DNA: a compilation of old and new data. J. Molec. Biol. **3**: 31-40, 1961a.

SUEOKA, N. Correlation between base composition of DNA and amino acid composition of protein. Proc. Nat. Acad. Sci. U. S. **47**: 1141-1149, 1961b.

SUEOKA, N. Compositional correlation between DNA and RNA. Cold. Spr. Harb. Sym. Quant. Biol. **26**: 35-43, 1962.

SUEOKA, N., and TS'AI-YING CHENG. Fractionation of nucleic acids with the methylated albumin column. J. Molec. Biol. **4**: 161-172, 1962.

TSUGITA, A. The protein mutants of TMV. J. Molec. Biol. **5**: 284-300, 1962.

TSUGITA, A., H. FRAENKEL-CONRAT, M. N. NIRENBERG, and J. H. MATTHAEI. Demonstration of the messenger role of viral RNA. Proc. Nat. Acad. Sci. U. S. **48**: 846-853, 1962.

VOLKIN, E., and L. ASTRACHAN. Phosphorus incorporation in *Escherichia coli* ribonucleic acid after infection with bacteriophage T2. Virology **2**: 149-161, 1956.

WALL, R. Overlapping genetic codes. Nature **193**: 1268-1270, 1962.

WATSON, J. D., and F. H. C. CRICK. Genetic implications of the structure of deoxyribonucleic acid. Nature **171**: 964-969, 1953.

WEISBLUM, B., S. BENZER, and R. W. HOLLEY. A physical basis for degeneracy in the amino acid code. Proc. Nat. Acad. Sci. U. S. **48**: 1449-1454, 1962.

WILSKA, A. P. N. Y. Times (Scientist derives way to see atom), Nov. 26, 1962.

WITTMAN, VON H. G. Ansätze zur Entschlüsselung des genetischen Codes. Naturwissen. **24**: 729-734, 1961.

WOESE, C. R. A nucleotide triplet code for amino acids. Biochem. Biophys. Res. Comm. **5**: 88-93, 1961.

WOESE, C. R. Nature of the biological code. Nature **194**: 1114-1115, 1962.

WOOD, W. B., and P. BERG. The effect of enzymatically synthesized ribonucleic acid or amino acid incorporation by a soluble protein-ribosome system from *Escherichia coli*. Proc. Nat. Acad. Sci. U. S. **48**: 94-104, 1962.

YAMAZAKI, H., and P. KAESBERG. Biophysical and biochemical properties of wild cucumber mosaic virus and of two related virus-like particles. Biochim. Biophys. Acta **51**: 9-18, 1961.

YANOFSKY, C., D. R. HELINSKI, and B. MALING. The effects of mutation on the composition and properties of the A protein of *Escherichia coli* tryptophan synthetase. Cold. Spr. Harb. Symp. Quant. Biol. **26**: 11-24, 1961.

YCAS, M. Correlation of viral RNA and protein composition. Nature **188**: 209-212, 1960.

YCAS, M. Replacement of amino acids in proteins. J. Theoret. Biol. **1**: 244-257, 1961.

YCAS, M. The coding hypothesis. Int. Rev. Cytol. **13**: 1-37, 1962.

YOSHIKAWA, H., and N. SUEOKA. Mechanism of chromosome replication in *B. subtilis*. Rec. Genet. Soc. Am. **32**: 127, 1962.

ZUBAY, G., and H. QUASTLER. An RNA-protein code based on replacement data. Proc. Nat. Acad. Sci. U. S. **48**: 461-471, 1962.