*THE NATURE OF THE BIOCHEMICAL CODE**

V. Yu. GAVRILOV and Yu. N. ZOGRAF

Usp. Fiz. Nauk 77, 597-620 (August, 1962)

RECENT years have marked very great advances in understanding the processes of protein synthesis in the living cell and in elucidating the role of specific biopolymers, the nucleic acids, in these processes. The greatest advances in this field have been the attainment under laboratory conditions of the biosynthesis of artificial analogs of the nucleic acids, the elucidation of the fundamental stages in protein synthesis, and the elucidation of the structures of the individual types of nucleic acids and the functions which they perform. Very recently, great advances have been made in studying the nature of the biochemical code, i.e., in finding out the laws relating the fine structural details of the nucleic acids to the arrangement of amino-acid residues in the proteins being synthesized by the cell.

This review has been written with the aim of acquainting the readership of physicists with the fundamental facts and theories in this swiftly developing field of natural science. Of necessity, we have had to write in a somewhat popularizing style, and cannot pretend at all to have presented the problems concerned in an exhaustive manner.

The cell contains two types of nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA); the latter consists of several fractions which differ essentially in their structure and functions. It is generally agreed at present (although objections are still heard from some individual specialists) that DNA occurs in the nuclei of cells, while RNA occurs both in the nuclei and in the cytoplasm. Many facts have been amassed proving that the DNA contains within itself the genetic information determining the transfer from generation to generation of the traits characteristic of a given cell or of an entire multicellular organism as a whole. Each of these facts taken separately might be questioned to some extent, but in combination, they leave no doubt as to the specific genetic role of the DNA in the cell.

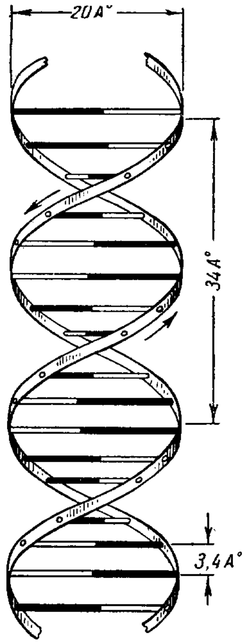

FIG. 1. Diagram of the two-stranded helical structure of DNA according to the Watson-Crick model.

The molecular dimensions are given for the B form of DNA (at high relative humidity).

In 1953 Watson and Crick analyzed the x-ray diffraction data on the sodium salt of deoxyribonucleic acid, and using a number of stereochemical considerations, they proposed their model for the structure of DNA.[1,2] According to their proposals, DNA is a double helix of diameter 18—20 Å and pitch 34 Å (Fig. 1), consisting of two sugar-phosphate chains (or strands) running in opposite directions, comprising phosphoric acid residues and molecules of the sugar deoxyribose. Molecules of purine and pyrimidine bases are attached to the sugar molecules, and project inside the double helix, being linked to each other by relatively weak hydrogen bonds (Fig. 2). At sufficiently high humidities, the planes containing the pairs of bases attached to the sugar molecules of the two oppositely-directed chains of the helix are perpendicular to the axis of the latter.

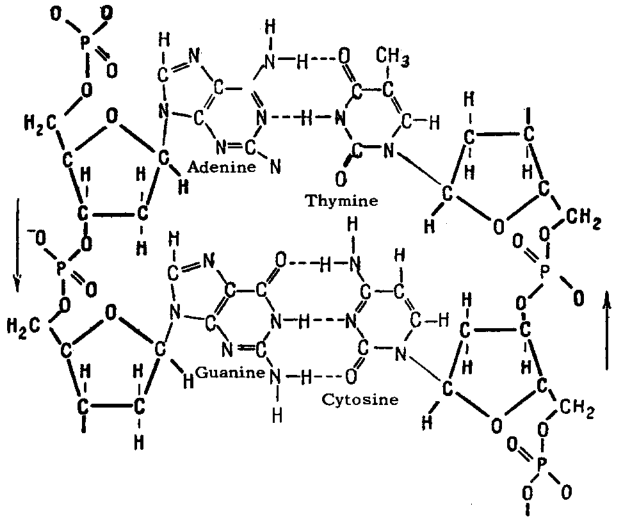

FIG. 2. Structure of a region in the DNA molecule.

The planes of the bases must be rotated perpendicular to the plane of the drawing. The sugar-phosphate chains are drawn in boldface. The dotted lines indicate the hydrogen bonds between the bases in the different strands.

*A paper presented at a seminar of the S. I. Vavilov Institute of Physics Problems on May 10, 1962.

The opposing directions of the sugar-phosphate chains of DNA result from the asymmetric structure of the sugar molecule; the helical structure of DNA becomes possible only under the condition that the two chains run in opposite directions. The fact that the two chains run in opposite directions has been established experimentally by biochemical methods.[3]

Deoxyribonucleic acid (DNA) contains four different bases: two purine bases, adenine (A) and guanine (G); and two pyrimidine bases, thymine (T) and cytosine (C). The fundamental link of DNA, consisting of a phosphoric acid residue, a sugar, and the base attached to it, is called a nucleotide. For stereochemical reasons, Watson and Crick concluded that the bases belong to the different strands of the molecule combine by hydrogen bonds in strictly defined pairs: A combines by two hydrogen bonds with T, and G by three hydrogen bonds with C. Such a model is in excellent agreement with the experimental data[4] and explains why in DNA the amount of A is always equal to the amount of T, and the amount of G equals that of C (the so-called Chargaff's Rule[5]). The composition of DNA differs considerably from species to species. For example, among microorganisms the ratio $(A+T)/(G+C)$ varies from 0.35 to 2.84. Among the higher plants, this ratio varies from 1.08 to 1.78. Among the invertebrates, it varies from 1.27 to 1.94, but among the vertebrates, only within the limits from 1.27 to 1.50.[6]

Besides the four bases mentioned, other bases are found in DNA. Among these, the most important are 5-methylcytosine, which replaces part of the cytosine in the DNA of a number of organisms, and 5-hydroxymethylcytosine, which replaces all the cytosine in certain phages. The sequence of nucleotides in DNA is not known,[7] but it is known to be unique for each species; in particular, the probability of finding two given nucleotides in adjacent positions differs for DNA of different species.[3] The molecular weight, and thus the length of the isolated DNA molecule, depends greatly on the method of isolation.[8] While ordinarily the molecular weight of DNA in isolated preparations varies from 5×10^6 to 15×10^6, recently DNA molecules obtained from T2 phage of length $52\,\mu$, corresponding to a molecular weight of about 110×10^6.[9] There are signs that, in certain physiological states of the cell, the DNA which it contains can separate at least partially into individual chains, i.e., transform from the double-stranded to the single-stranded state.[10] We must note that upon breaking the hydrogen bonds, the two strands of the DNA helix cannot be simply detached from each other, and the mechanism of untwisting of two such strands is not at all clear as yet. Most organisms contain double-stranded DNA. However, there are exceptions. Thus, e.g., the phage ϕX174 contains single-stranded DNA.[11] Such forms of DNA, naturally, do not follow Chargaff's Rule.

The rule of complementarity of bases of the different strands of DNA, i.e., the formation of hydrogen bonds only between certain definite pairs of complementary bases (between A and T, and between G and C), has the result that the sequence of bases in one strand of the DNA completely determines that in the other strand, which is complementary to it. On this basis, Watson and Crick[12,13] proposed that DNA can reproduce itself by reduplication, in which each of the strands of the original molecule serves as a template for the synthesis of one of the two daughter molecules of DNA. Here, as the weak hydrogen bonds between the bases are broken, the strands are separated, and then each of the liberated bases of the strand interacts with a complementary base of a free nucleotide so as to bring about polymerization of a complementary strand having a specific sequence of nucleotides. This property of DNA has made it possible to suggest that this is just how the genetic information contained in the DNA molecule is duplicated exactly and transferred from generation to generation in the course of cell division. Meselson and Stahl[14] have studied the distribution of isotopic labels introduced into DNA during the reproduction of the DNA in bacterial cells, and showed that the new DNA molecules contain one old and one newly-synthesized strand. That is, DNA synthesis follows the Watson-Crick scheme of reduplication. They showed as well that the strands separate when the isolated DNA is heated to about 100°C (denaturation). More recently, Doty[16-18] has found that under certain conditions of cooling, these DNA strands which have been separated by heating combine again into a double helix (the so-called renaturation of DNA). Doty also obtained biologically-active ''hybrid'' DNA molecules. Here the DNA was taken from two related species of bacteria, artificially denatured, and then reconstructed into two-stranded molecules.

A very important advance in biochemistry was the extracellular biosynthesis of the DNA molecule achieved by Kornberg in 1958 from nucleoside triphosphates, with the aid of a specific enzyme, DNA polymerase, which he had isolated from bacteria.[19-22] Here, regardless of the organisms from whose cells this enzyme has been isolated, the synthesized DNA resembles that which was added to the system as a primer. By using this enzyme without adding a DNA primer to the system, it was possible to obtain artificial analogs of DNA: two-stranded synthetic A-T and G-C copolymers. Here the poly-AT consisted of two strands, in each of which A and T alternate regularly, while the poly-GC contained one strand consisting entirely of G and another consisting entirely of C. It was shown that the synthesis of DNA in this system takes place on one strand of the added primer, which constructs a strand complementary to itself. If we use a single-stranded DNA as the primer, the synthesized DNA will be double-stranded. Hence we may assume that the processes of separation of the strands and of polymerization of a complementary strand are distinct, but the concrete physicochemical mechanism of neither of these is clear.

Both the nucleus and the cytoplasm contain a large amount of the other nucleic acid, ribonucleic acid (RNA). It differs chemically from DNA in containing a different sugar (ribose), and instead of thymine, it contains uracil (U). The cells contain several different types of RNA, low-molecular-weight soluble RNA (sRNA), consisting of about 50—100 nucleotides, and highly-polymerized RNA having molecular weights of the order of millions. In spite of the fact that these molecules are probably single-stranded, they possess a certain degree of order. It is assumed that they contain loop-like helical regions, in which the nucleotides are linked by hydrogen bonds.[23,24]

The specific properties of each cell is determined primarily by the composition of its proteins. In particular, all the enzymes of the cell are proteins. The number of types of different proteins in the cells is very large, and even in small bacterial cells the number exceeds several hundred. The proteins are polypeptide chains consisting of several hundred, or sometimes of several tens of amino-acid residues linked together by peptide bonds (Fig. 3). They have a complex secondary structure (the configuration of the polypeptide chain, which is determined by the system of hydrogen bonds between the amino-acid residues) and tertiary structure (the spatial arrangement of the chains, determined mainly by disulfide bonds between the residues of the sulfur-containing amino-acid cysteine). Many proteins consist of several separate polypeptide chains joined by disulfide bridges. The polypeptide chain of a protein has a definite directional character, since the peptide bond is formed between C and N atoms of adjacent amino-acids, i.e., unsymmetrically. A number of authors, in particular Crick,[25] have advanced the hypothesis, which is confirmed by experimental data,[26] that the primary structure (i.e., the sequence of amino-acid residues in the polypeptide chain of a protein) determines its secondary and tertiary structure, and hence, its functional properties. The bulk of the proteins is synthesized in the cytoplasm, the synthesis fundamentally taking place in special structured bodies in the cell, which are granules of ribonucleoprotein, or the so-called ribosomes. The ribosomes are very rich in RNA and contain RNA of two types, having molecular weights of $5-6 \times 10^5$ and $1.1-1.3 \times 10^6$.[27,28] The structure of the ribosomes is not known as yet. When the concentration of Mg^{++} in the medium is small, two types of ribosomes are found, those with a sedimentation constant 30 S and molecular weight of about 0.95×10^6, and those with sedimentation constant 50 S and molecular weight 1.85×10^6. When the Mg^{++} concentration is increased, one observes the aggregation of these ribosomes into particles having higher sedimentation constants, the 70 S and 100 S ribosomes:[29]

$$2(30s)+2(50s)\rightleftarrows 2(70s)\rightleftarrows(100s).$$

Protein synthesis takes place in the 70 S-ribosomes, and apparently in the 100 S-ribosomes, while the other configurations of ribosome particles are not active.

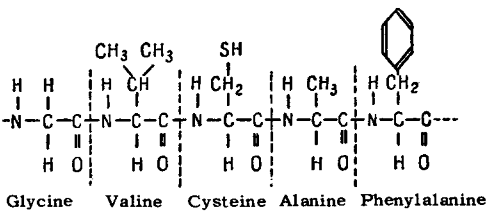

FIG. 3. A region in a protein chain.
The peptide bonds are formed between the C and N atoms of adjacent amino-acid residues, which are separated by the dotted lines.

Beginning in 1954-1956, a vast study has been made to elucidate the entire system of protein synthesis. It has been based on the studies of the group of Zamecnik and Hoagland,[30-34] who have shown that this process takes place in several stages. To start, under the action of certain enzymes, the amino-acids become activated to form specific compounds, the aminoacyladenylates. Here, a necessary participant in the process is the compound adenosine triphosphate (ATP), which possesses high-energy bonds. During this process, inorganic pyrophosphate splits out of the ATP, and the remaining part of the ATP combines with the amino-acid. In the next stage, the activated amino-acid combines with a molecule of sRNA with the liberation of adenosine monophosphate. Here the amino-acid is linked to a certain definite end of the sRNA; all of the sRNA molecules have the same trinucleotide sequence (C—C—A) at this end. These first two stages involve specific activating enzymes. Here the sRNA molecules are evidently specific with respect to the amino-acids, but there are indications that the list of sRNA molecules is the same, even among the most highly differing organisms. The activating enzymes are also specific for the amino-acids. Apparently, each amino-acid has its own sRNA and its own activating enzyme. The last stage of this process, which takes place within the ribosome, is the arrangement of the sRNA molecules with their attached amino-acids on the template determining the structure of the protein, and the linking of the amino-acids by peptide bonds, i.e., the direct formation of the polypeptide chain of the protein. Here the sRNA is liberated, and goes back into solution. The concrete mechanism of this stage is not yet clear. Such phenomena of synthesis of polymer chains on templates are not yet known in chemistry at present, and none of the stages of this type of synthesis have been studied. It is interesting to note that the linkage of amino-acids into the peptide chain takes place only in a certain direction sequentially, and only the completely formed polypeptide chain of the protein is detached from the template. Thus, the synthesis of the peptide chain of hemoglobin begins at the N-terminal[35] and takes 1.5 minutes to complete. That is, approximately two amino-acids are polymerized every second.[36]

For a long time, a number of investigators have expressed the opinion that the structure of the nucleic acids determines the properties of the proteins synthesized by the cell. For example, a very convincing experiment upon which this viewpoint has been based was that conducted by Schramm and Schuster on tobacco mosaic virus.[37]

The viruses, which parasitize plant or animal cells, always contain a certain amount of proteins and a nucleic acid. As a rule, this is DNA in animal viruses and RNA in plant viruses. By themselves, viruses are not capable of multiplication, and the process of multiplication takes place only within cells infected with them. When a cell is infected with a virus, the synthesis begins therein of the nucleic acid characteristic of the virus and of a set of specific proteins, which either enter directly into the composition of the new virus particles, or are necessary as enzymes regulating the processes of synthesis of biochemical compounds which are new for the given cell.

Tobacco mosaic virus, which attacks tobacco leaves and causes a characteristic disease therein, is structurally the simplest of the viruses thus far studied. It contains only one molecule of RNA, consisting of about 6500 nucleotides, and 2200 protein molecules. The latter are all identical, each consisting of a polypeptide chain containing 158 amino-acid residues, whose sequence has been completely determined.[38] By treating the virus particles with nitrous acid, Schramm and Schuster showed that this compound is a chemical mutagen, i.e., it can create in the virus new properties, which are transmitted from generation to generation. As is known, nitrous acid acts on the purine and pyrimidine bases in RNA, transforming adenine into guanine or cytosine into uracil at certain points in the molecule. On the other hand, it does not act directly on the proteins. By reaction of nitrous acid with the RNA of tobacco mosaic virus, Tsugita and Fraenkel-Conrat[59] have discovered an entire set of mutants distinguished by the replacement of some amino-acid residue by another in the peptide chain of the protein molecules of the virus. That is, a change in the RNA led to a change in the amino-acid residues forming the protein.

As the theories on the stages of protein synthesis in the cell developed, the conception of a connection between the composition of the proteins and the base sequence in the nucleic acid carrying the genetic information encountered a number of difficulties. No direct connection was visible between the DNA in the nucleus of the cell and the proteins being synthesized in the ribosomes. The highly-polymerized RNA contained within the ribosomes may differ completely in its overall base composition from that of the DNA, and varies exceedingly little from cell to cell.*

This difficulty was eliminated in 1960 with the discovery of a special type of highly-polymerized information-RNA or "messenger-RNA" (mRNA), which is synthesized in the nucleus using the DNA as a template, and then emerges into the cytoplasm and takes part as a template in protein synthesis in the ribosomes. This mRNA amounts to only an insignificant fraction of the total RNA of the cell, usually no more than 2—3%. It differs from the rest of the RNA in that it has a short lifetime, of the order of several minutes. Supposedly, each molecule of this type of RNA brings about the synthesis in the ribosome of one protein molecule, and then decomposes. Upon pulsed introduction into the cell of labeled nucleotides, practically all of the label appears precisely in the mRNA. By this method, synthesis of mRNA has also been demonstrated in bacterial cells infected by a phage; here its nucleotide composition turned out to correspond to that of the DNA of the phage.[39-41] This has been done since on several other objects as well.[42,43] It turned out that this RNA is incorporated into the 70 S and 100 S-ribosomes[43], about 10% of the ribosomes in the cell containing it at any one time.[44] It is the mRNA in particular, rather than the ribosome particles themselves, which determines the specificity of the synthesized protein. Indeed, in phage infection of bacteria, the new specific proteins determined by the phage DNA are synthesized in the old bacterial ribosomes existing prior to the infection. However, the synthesis involves new mRNA corresponding in nucleotide composition to the DNA of the phage, rather than to the DNA of the bacterium.[45] This mRNA has a molecular weight of about 3×10^5.

Not only does the total base composition of the mRNA correspond to that of the DNA, but the mRNA also has a base sequence complementary to that of the DNA. This can be shown by combined renaturation in vitro of mRNA with the corresponding DNA which has been subjected to preliminary separation of the strands by heating. Here "hybrid" complexes are obtained, consisting of one strand of DNA and one strand of mRNA.[46] Complexes consisting of DNA and mRNA have also been isolated directly from cells,[47-48] although indeed their structures have not been determined yet.

Enzyme systems have been isolated from various organisms, which bring about RNA synthesis involving DNA as a template.[49-53] The nucleotide composition and sequence in this artificially synthesized mRNA completely corresponds to the composition and sequence of the primer DNA. Here the RNA synthesis

*Nevertheless, Belozerskiĭ and Spirin have shown a certain correlation between the nucleotide compositions of the DNA and the total RNA in microorganisms (A. N. Belozerskiĭ and A. S. Spirin, Nature 182, 111 (1958)).

may be carried out using either double-stranded or single-stranded DNA as primer.

Thus it seems that the template for mRNA synthesis may be either of the DNA strands, and two strands of mRNA are simultaneously synthesized, complementary to both of the DNA strands.[53]

The assumption of a relation between the structure of the DNA and that of the proteins synthesized by the cell led Gamow[54,55] in 1954 to formulate the problem of a biochemical code, i.e., the problem of how the base sequence in the nucleic acid (DNA or the RNA determined by it) determines the sequence of amino-acids in the protein. Here Gamow still made the assumption that the proteins are directly synthesized by the DNA. Of all the amino-acids occurring in nature, only twenty occur in the proteins of all organisms; these are listed in Table I.

Table I. The amino-acids directly taking part in forming the polypeptide chain of a protein

Amino-acid	Symbol	Amino-acid	Symbol
Alanine	Ala	Leucine	Leu
Arginine	Arg	Isoleucine	Ileu
Aspartic acid	Asp	Lysine	Lys
Asparagine	AspN	Methionine	Met
	(or $AspNH_2$)	Proline	Pro
Valine	Val	Serine	Ser
Histidine	His	Tyrosine	Tyr
Glycine	Gly	Threonine	Thr
Glutamic acid	Glu	Tryptophan	Try
Glutamine	GluN	Phenylalanine	Phe
	(or $GluNH_2$)	Cysteine	CySH

The other amino-acids occurring in nature either are not incorporated into proteins at all, or they are found in only a small number of specific proteins (hydroxyproline in collagen, tyrosine-0-sulfate in fibrinogen, etc.). Crick was the first to propose that these "irregular" amino-acids are formed by modification of certain of the corresponding twenty "regular" amino-acids after the latter have already been incorporated into the polypeptide chain of the protein, which has been built up from only the twenty "regular" amino-acids. Thus, the sequence of the four different bases in the nucleic acid must determine the sequence of the twenty different amino-acid residues in the protein. Since DNA (or RNA) contains four different bases, in a nucleic acid the number of differing coding units consisting of n nucleotides is 4^n, if we impose no limitations on the form of the coding units. The necessity of coding the twenty amino-acids with these coding units implies that the inequality $4^n \geq 20$ must be satisfied. Hence, the coding units in the nucleic acid must consist of at least three nucleotides (a triplet), if, of course, the dimensions of the coding units are the same for all amino-acids.

In general, the coding units in the nucleic acid determining consecutive amino-acids in the polypeptide chain of the protein might overlap. That is, they might have one or several nucleotides in common. If the consecutive coding units in the nucleic acid overlap, we will observe a correlation between the consecutive amino-acids in the protein. Such a correlation could be weak only in case that the dimensions of the overlap are much smaller than those of the coding unit itself. The problem of the correlation of consecutive amino-acids in proteins has been investigated by Yčas, Gamow, and Rich,[56,57] who made an analysis of the distribution law of consecutive amino-acid residues for a number of different proteins for which the amino-acid sequence was known for at least part of the polypeptide chain. They showed that no correlation is observed between consecutive residues. Their conclusions were later confirmed by Brenner,[58] and it may be considered to be established that if the code is a triplet code and is universal (i.e., the same for all organisms), overlaps between adjacent triplets are ruled out.

The lack of overlap between consecutive coding units is also demonstrated by an analysis of the amino-acid sequence in the chain of the tobacco mosaic virus protein, when changed upon treatment of the virus RNA with mutagenic agents. This analysis was carried out in 1960–61 by Fraenkel-Conrat and Tsugita[59,60] and by Wittmann,[61] and showed that in most cases the mutation leads to the replacement of only one of the amino-acids in the polypeptide chain of the protein, without affecting its neighbors. Even in the cases in which two amino-acid residues in the protein had been replaced simultaneously, they were found to be located in different parts of the molecule.

The universality of the code is indicated by the probable nonspecificity with respect to species of the sRNA taking part in protein synthesis. In addition, the universality of the code is indicated by a study by Sueoka [62] on the correlation between the amino-acid composition of the total proteins and the nucleotide composition of the DNA in various bacteria. The correlation found for a number of amino-acids turned out to be the same for all eleven species of bacteria studied.

The abundance of the amino-acids in proteins differs considerably from a random distribution.[63,64] Certain amino-acids, e.g., leucine or alanine, occur about ten times more abundantly than others such as cysteine or methionine. This predominance of certain amino-acids over others can arise either from the fact that certain coding units in the nucleic acid are more frequent than others, as is known, or by virtue of correspondence of single amino-acids to a larger number of different coding units. In the latter case, in which one amino-acid may correspond to several different coding units, the code is said to be degenerate.

The idea that the amino-acid sequence in the protein is determined by the base sequence in the DNA directly elicited the proposal of a number of concrete coding systems. In particular, Gamow's studies[54-56,64] pointed out the fact that all 64 possible triplets may be divided into 20 groups, each of which corresponds to one "regular" amino-acid, and differs from the other groups in nucleotide composition. In such a case, one amino-acid may correspond to several triplets which differ only in the order of the nucleotides which they contain. In addition, a number of other coding systems have been proposed.[56,65] In none of these systems was it proposed that meaningless nucleotide combinations might exist, not corresponding to any amino-acid.

Almost immediately after the appearance of the first theories on the biochemical code, difficulties arose involving the fact that, if the code is non-overlapping, we must still assign a method of reading the message, since there is no unequivocal interpretation of the recorded information. Indeed, we may assume that the information is recorded in a single-stranded sequence of nucleotides (e.g., as in mRNA) in which there is, naturally, a defined direction:

$$\ldots 123456789 \ldots$$

Then, provided that the code is non-overlapping, in general this record can be read in several manners, e.g., in three manners for a triplet code:

$$\ldots 123-456-789 \ldots,$$
$$\ldots 1-234-567-89 \ldots,$$
$$\ldots 12-345-678-9 \ldots$$

We can obtain an unequivocal result either if there are special "commas" between the triplets, or if there is a starting point for reading and interpreting the message, and we proceed from this starting point triplet by triplet in sequence. Yčas[66] has even proposed that all three messages may be read, as obtained from the given message by shifting, and that the protein is formed from the three polypeptide chains obtained by reading all three of the messages coded in the single sequence of nucleotides. However, this is clearly impossible, since in many proteins the number of amino-acid residues present is not divisible by three.

In order to obtain an unequivocal result in reading the information, Crick, Griffith, and Orgel[67] have proposed a so-called "comma-less" code. In a code of this sort, the amino-acids correspond only to certain "meaningful" triplets, such as cannot give other meaningful triplets by overlap. For example, a triplet of the type AAA cannot be meaningful, since an overlap of two such triplets gives the same triplet AAA. From the four different bases, we can by various methods construct just twenty triplets satisfying this condition. In such a code, the information can be read simultaneously throughout the nucleic acid molecule, rather than sequentially from the end. Further study has not confirmed this system.

Further difficulties in the problem of unequivocal reading involve the two-stranded structure. The point is that the two-stranded structure does not have a preferred direction, since the strands run in opposite directions. Hence, if the direction is assigned for reading the information from the mRNA onto the protein, and we can eliminate the ambiguity by assigning a starting point, then to read the information from the DNA, we need to distinguish the direction of reading. If there is no preferred direction for reading the information from the DNA, unequivocality may be attained either if the information in both strands is identical,[68] or if one of the strands lacks information.[69] However, if there is a starting point in the single-stranded sequence (mRNA), such that the information can be read unequivocally, there must also be such a starting point in the DNA. Such a starting point in the DNA may distinguish the strand from which we must read the information, at the same time distinguishing the direction of reading as well, and marking off the regions in the DNA responsible for the synthesis of different polypeptide chains.

All of the difficulties mentioned above have had the result that no clear system of coding has appeared, in essence, having convincing advantages over the others, in spite of the considerable number of theoretical studies concerned with the problem of the biochemical code since 1954. Even the hypotheses of the existence of this sort of code were lacking in experimental proof. The first direct experimental proof of this sort was reported at the Fifth International Biochemical Congress, which took place in Moscow in August, 1961, a study by Nirenberg and Matthaei.[70,71] They isolated a ribosome fraction from intestinal bacteria (Escherichia coli) by centrifugation, and mixed it with the supernatant liquid, which contained, in particular, sRNA and the activating enzymes. They introduced into this system amino-acids and polyribouridylic acid (poly-U), which was obtained by Ochoa's method, and which contains as a base only uracil. Thereupon they observed the appearance of polypeptides consisting of only one type of amino-acid, or polyphenylalanine. They showed later that the process of synthesis of polyphenylalanine proceeds through all the usual stages of protein synthesis: phenylalanine is activated, then it combines with the appropriate sRNA, and then polyphenylalanine is synthesized in the ribosomes with poly-U acting as the template.[72] At the same time, they demonstrated also the single-stranded character of the template in protein synthesis in the ribosomes, since when they added to the system polyadenylic acid (poly-A), which forms a two-stranded helical complex with poly-U, the synthesis of polyphenylalanine was suppressed.

The study of Nirenberg and Matthaei indicates that the specific group of nucleotides in mRNA coding phenylalanine consists of UUU... This rules out the "comma-less" code of Crick, Griffith, and Orgel. The fact that such a combination turned out to be meaningful compelled them to assume that unequivocality of reading of the information on the nucleic acid must depend on special points determining where to start reading, and the reading must proceed from these points sequentially by whole coding units of nucleotides. The evidence for the presence of a special starting point permitted Crick to formulate a hypothesis on the type of code. He proposed that DNA contains starting points for reading, which divide the regions in the DNA responsible for the synthesis of different polypeptide chains, and that the reading proceeds sequentially from the starting point by whole coding units of nucleotides. Here the size of the coding units of nucleotides is the same for all the amino-acids. He also proposed that the code is degenerate, i.e., that one amino-acid can correspond to several different coding units of nucleotides, and that meaningless combinations of nucleotides exist, which do not correspond to any amino-acid. In order to test this hypothesis, he studied, together with Brenner, Barnett, and Watts-Tobin,[73] the mutations produced by acridines in a special region of the DNA of T4 bacteriophage. The bacteriophages, or bacterial viruses, contain one large DNA molecule and a small number of soluble proteins within a protein envelope. When a bacterium is infected by a phage (the T-even phages attack E. coli), all of the DNA of the phage but only an insignificant fraction of its proteins penetrate the bacterium. After the infection, an intensive synthesis begins in the bacterium of specific RNA and proteins, and in six or seven minutes, the synthesis of new phage DNA begins. About a half-hour after the infection, about 100 mature phage particles have accumulated in the bacterium (as many as several hundred with certain phages). The cell membrane of the bacterium disintegrates (the so-called lysis of the bacterium), and the phages which emerge are capable of infecting new bacteria. If one bacterium is infected simultaneously by two phages which differ in certain genetic traits, then recombination of the genetic material of these phages takes place within the bacterium. Here the DNA molecules of the two phages are broken at the same point, and part of the molecule of one phage is spliced onto the portion of the DNA molecule of the other phage which supplements it to form a whole molecule. It seems that such a rupture can take place in any region of the DNA molecule. Among the descendents of the original phages are found phages possessing mixed genetic material, and combining the corresponding traits of both the original phages. Obviously, in this sort of process, the probability that two closely-spaced genetic regions in the DNA molecule of one of the original phages will be separated and end up in the DNA molecules of different newly-formed phages will be lower than for two distant regions.

The experimental study of the frequencies of recombinations of certain types and the comparison of the frequencies with each other makes it possible to construct rather detailed genetic maps of the phages. These maps indicate the sequence along the DNA molecule of the phage in which the genetic regions responsible for the appearance of certain traits occur, and they give the distances between these regions in provisional units of the reciprocal of the frequency of a recombination of the indicated type.

If the phage is plated out at high dilution on a bacterial culture plate, the descendants of each individual phage will "eat away" the bacteria, i.e., lyse them, thus producing "empty" spots or "plaques" in the bacterial culture. T4 phage of the standard "wild" type grows on various strains of E. coli, in particular strains B and K12. A phage in which a certain region of the genetic material, the rII locus, has been injured does not grow on strain K12, but forms r-type plaques on strain B; these differ from the plaques of wild-type phage in being larger. Benzer[74] has shown that the rII locus consists of two independently functioning regions, cistrons A and B, which are responsible for the synthesis of two different proteins. However, both of the proteins are responsible for the same trait, since damage to either of these two cistrons leads to r-type. It had been assumed even prior to this[75] that when DNA is treated with specific chemical compounds, the acridines, the incorporation of the acridine into the DNA structure leads in a subsequent reduplication either to deletion of one of the nucleotides or insertion of a nucleotide into the polynucleotide chain. This does not exclude the possibility of simultaneous deletion or insertion of several nucleotides. Crick has pointed out that if his proposals as to the type of code are true, the deletion of a nucleotide must result in a shift in the reading of all the coding units by one base to the right from this site on (Fig. 4a). The insertion of a nucleotide must correspondingly result in a shift of one base to the left in the reading (Fig. 4b). The effect of this must be that the sequence of amino-acids in the protein determined by this cistron is completely changed at and beyond that amino-acid determined by the coding unit of nucleotides which has undergone insertion or deletion. Here the amino-acid sequence is changed in different ways, depending on whether the shift was to the left or the right. The protein obtained is essentially different from that of the wild-type phage, and this is what is involved in the change in behavior of the mutant phage. If there are meaningless combinations of nucleotides, the possibility is not excluded that such a meaningless group may appear at some point in the DNA when a certain shift has taken place. The appearance of a meaningless group of nucleotides must lead to a break in the polypeptide chain of the protein at this point.

The simultaneous presence of two mutations, one of which involves the deletion of a nucleotide and the other an addition at any other site on the polynucleotide chain of the DNA in the same cistron, according to Crick's

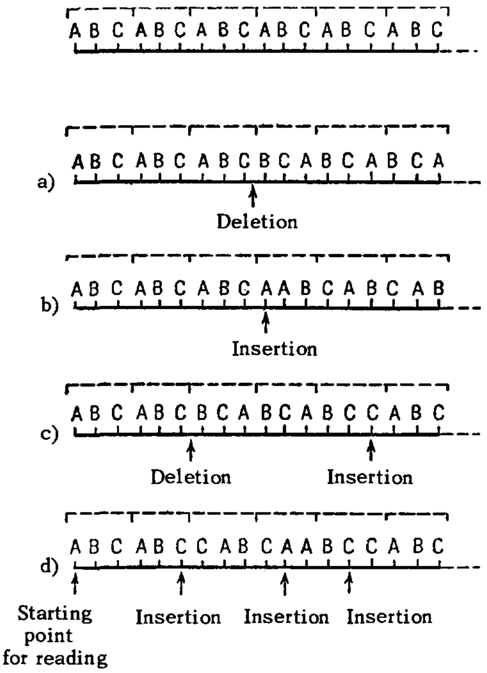
A B C A B C A B C A B C A B C A B C
a)
A B C A B C A B C B C A B C A B C A
Deletion
b)
A B C A B C A B C A A B C A B C A B
Insertion
c)
A B C A B C B C A B C A B C C A B C
Deletion
Insertion
d)
A B C A B C C A B C A A B C C A B C
Starting point for reading
Insertion
Insertion
Insertion

FIG. 4. The effect of insertion and deletion of bases in the nucleic acid on the reading of the information.

The letters A, B, and C indicate the first, second, and third nucleotides of the triplet, and may be different nucleotides. The code is assumed to be a triplet code. The reading proceeds from left to right in whole triplets.

conception must lead to a shift in the reading one base to the right from the first mutation on, and one base to the left from the second mutation on. Obviously, in such a case the information being read will be altered only in the region between these mutations (Fig. 4c). Thus the amino-acid sequence will be altered only between the amino-acids determined by the mutated units of the DNA, while the remainder of the polypeptide chain of the protein of such a mutant phage must coincide with that of the wild-type phage. If the presence of such an altered region in the polypeptide chain has no essential influence on the function of the protein, the appearance of the second mutation restoring the amino-acid sequence in the protein from some certain point on will result in a restoration of the phage to a form differing little from the wild type, or a so-called pseudo-wild type.

Indeed, Crick, Brenner, Barnett, and Watts-Tobin[73] have found inverse mutations in the first quarter of cistron B of the rII locus of T4 phage. In the presence of a direct acridine mutation, they restore the phage to a pseudo-wild type, while they do not coincide in position on the genetic map with the direct mutation. Such inverse mutations are called suppressors, since their appearance suppresses the action of the direct mutation. The presence of suppressors has also been found in the hIII locus of the same T4 phage by Jinks.[76] By using a mutation which they call FCO, produced by the action of one of the acridines (proflavin), Crick and his associates have found 18 spontaneously-appearing suppressors for this mutation. They are distributed in a region occupying about one-tenth of the cistron. Here we must bear in mind that each of these mutations taken by itself, i.e., obtained not as an inverse mutation of FC O but directly from the wild-type phage, converts the phage again to the r-type, in complete agreement with Crick's ideas. For each such mutation taken by itself, suppressors may arise, which restore the phage again to pseudo-wild type. Such mutations are hence suppressors of suppressors of the mutation FC O. Crick and his associates have obtained an entire series of mutations of this type, both spontaneously and by the action of acridines (Fig. 5). In addition, they have found an entire series of mutations suppressing the action of the suppressors of suppressors taken by themselves (i.e., in other words, they have obtained suppressors of suppressors of suppressors of FC O) (Fig. 5).

Let us suppose that mutation FC O consists in the insertion of one nucleotide into the polynucleotide sequence of the phage DNA. Then the suppressor mutations must consist in the deletion of one nucleotide, while the suppressors of suppressors again consist in the insertion of one nucleotide, etc. It follows from Crick's hypothesis that all double mutants in which both mutations have involved deletion (or insertion) of nucleotides must be of r-type. Indeed, all of the double mutations in which the mutation FC O has been linked with any of its suppressors of suppressors by recombination are of r-type. The double mutants in which two different suppressors of mutation FC O are linked are also of r-type, as well as those which contain two suppressors of different suppressors of mutation FC O.

The limited size of the region in which the suppressors of a given mutation occur may be explained by the existence of meaningless combinations of nucleotides. Indeed, if a meaningless combination, not corresponding to any amino-acid, arises from the shift in reading between the deleted and inserted bases, the polypeptide chain of the protein will thus be interrupted, and a whole, functionally-active protein molecule will not be synthesized. Hence, the suppressors of a given mutation may be located only within a certain limited region of the genetic material around this mutation. Correspondingly, according to Crick's hypothesis, not every mutation involving deletion of a base can be a suppressor of a mutation involving insertion of a base. Restoration of the phage to wild or pseudo-wild type can be observed only for those double mutants in which no meaningless combination of nucleotides appears in the region of shifted reading. Naturally, the size of such a region might be quite different for right and left shifts. When the double mutation is a combination of mutations, one of which (provisionally) corresponds to the insertion of a nucleotide in the DNA molecule, and the other to a deletion, such a mutant will give rise

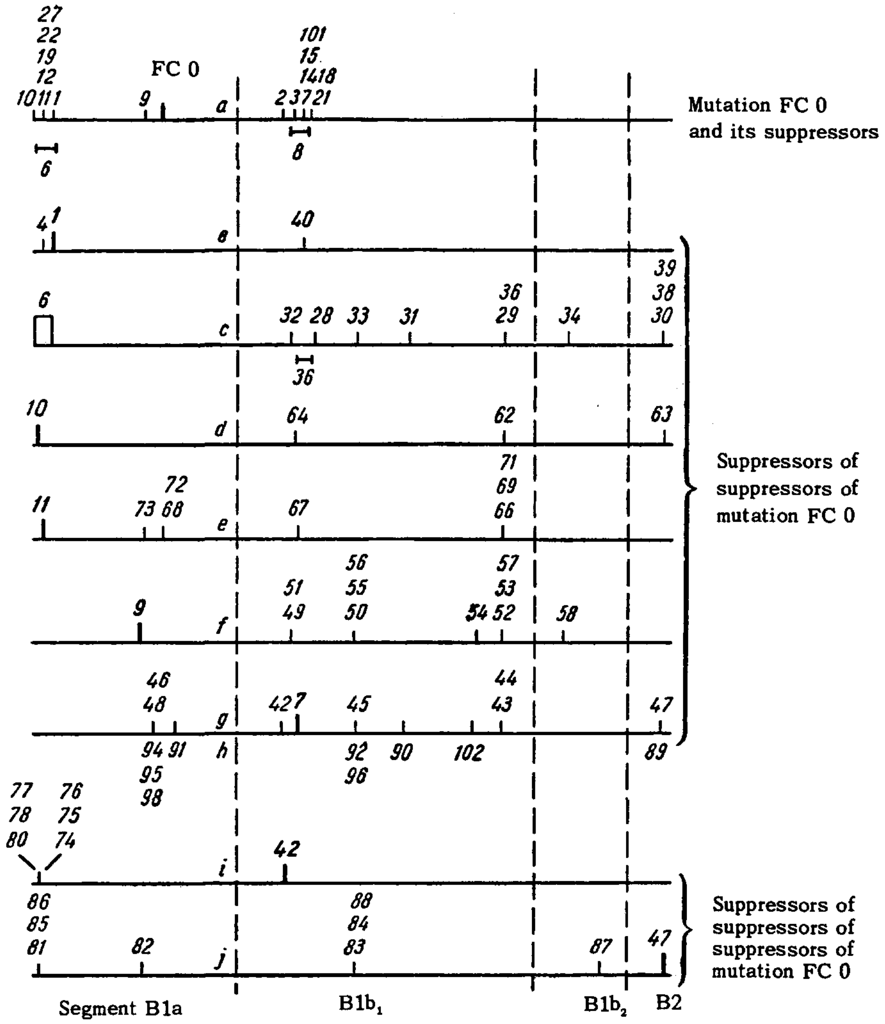

FIG. 5. Map of the left end of cistron B of the rII locus of T4 phage, showing the positions of the mutations studied by Crick, Brenner, Barnett, and Watts-Tobin.

Each line shows suppressors of the mutation labeled in boldface. All of the mutations were obtained spontaneously except for those on line h, which were obtained by the action of acridine.

Table II. The types exhibited by double mutants

Mutants crossed → ↓	41	FC 0	40	42	58	63	38
1	$\underline{w}$	$\underline{w}$	$\underline{w}$		w		w
86		w	w	w	w	w	
9	r	$\underline{w}$	w	w	$\underline{w}$		w
82	r		w	w	w	w	
21	r	$\underline{w}$			w		w
88	r	r			w	w	
87	r	r	r	r			w

The table gives the types exhibited by the double mutants obtained by recombination of the mutants of opposite sign indicated by the rows and columns, w = wild or pseudo-wild type; r = r-type. The mutants which were used for the isolation of the suppressors are underlined.

to pseudo-wild type, provided that no meaningless combinations appear between these mutations when the starting point for reading is shifted. The result will be r-type if a meaningless combination appears between the mutations. Twenty-eight double mutants were tested, combining mutations of opposite signs, and it was found that nineteen of them were of pseudo-wild type, and nine cases gave r-type (Table II). Again, this confirms beautifully Crick's assumptions on the character of the code.

According to Crick's hypothesis, if all the amino-acids correspond to groups of n nucleotides each in the DNA molecule, then when just n nucleotides have been deleted from (or added to) the nucleic acid molecule, even if located at different sites in the cistron, the reading of the information is shifted only between the first and the n-th deleted nucleotides, whereas the original reading is restored beyond the last deleted nucleotide (Fig. 4d). Hence, if no meaningless combinations appear in the region of shifted reading, then such an n-fold mutant must necessarily be of pseudo-wild type. Six of such triple mutants were studied, combining three mutations of the same sign. All these triple mutants were of wild or pseudo-wild type. Thus it was shown that the code is a triplet code, provided, of course, that the action of proflavin actually deletes or inserts a single base in every case, rather than several in sequence.

However, if the mutation results from the deletion or insertion of several bases in the DNA, the number of nucleotides in the coding unit is larger, but is a multiple of three. However, the latter case is hardly likely.

According to Crick's theory, if the starting point for reading is deleted from a given cistron, what will happen, so to speak, is the merging of two regions in the DNA governing the synthesis of two different protein molecules. Consequently, one molecule instead of two will be synthesized. This has been tested by Crick et al.[73] on a deletion (loss of an entire section of the genetic material) partially covering both cistrons A and B of the rII locus. Ordinarily, these cistrons function independently, and injury to one of them does not affect the function of the other. It was found that an acridine mutation in the left end of cistron A, provided that they are combined by such a deletion, interferes with the function of cistron B. That is, when such a deletion is present, these cistrons turn out to be merged.

Thus the genetic experiments described above have fully confirmed Crick's hypothesis on the nature of the code. The code has actually turned out to be a triplet code (or at least, the length of the coding unit is a multiple of three nucleotides), and is non-overlapping and degenerate. We can see that it is degenerate from the fact that the length of the region in the cistron encompassing the suppressors of a given mutation amounts in certain individual cases to several tens of triplet units. Hence, in spite of the fact that the starting point of reading was shifted when the original mutation occurred, and the meaning of the triplets following this shift became completely different from that in the unmutated molecule, often throughout several tens of triplets, we do not find a single meaningless one. Such a situation can occur only if the code is considerably degenerate.

The classic experiment of Crick, Barnett, Brenner, and Watts-Tobin has revealed the nature of the code, and in particular, it has confirmed the existence in the genome of special points determining the starting point for reading of each genetic "sentence" corresponding to one of the proteins being synthesized. However, this sort of experiment cannot in principle provide us with the "dictionary" itself, i.e., determine which concrete combination of the four possible bases corresponds to each of the amino-acid residues in the protein chain. This problem, or at least a considerable part of it, has been solved by the biochemists.

The nucleotide composition of the triplets can be found by studying the incorporation of amino-acids into protein in a system of ribosomes and supernatant liquid, when polyribonucleotides of various compositions are added to the system. The polyribonucleotides are synthesized from ribonucleoside diphosphates, using the enzyme polynucleotide phosphorylase, which was first isolated by Ochoa.[77] The nucleotide composition of the polyribonucleotides thus synthesized does not depend on the nucleotide composition of the primer, which in this case can be any fragments of polyribonucleotides or even dinucleotides. Rather, the nucleotide composition is determined solely by the relative concentrations of the various ribonucleoside diphosphates in the medium. Here the nucleotide sequence in the chain of the polyribonucleotide thus synthesized is completely random. If we provide the Ochoa enzyme with only one type of ribonucleoside diphosphate, a polynucleotide is formed containing only one type of base, e.g., poly-U. Analogously, in the Ochoa system we may also obtain polynucleotides containing only two or three types of base, e.g., poly-UC or poly-UGC, etc.

In late 1961 and early 1962, Ochoa[78-81] and concurrently Nirenberg and Matthaei[82] conducted a series of studies on the composition of the polypeptides formed in a system of ribosomes and supernatant liquid upon addition of various polyribonucleotides. In studies on the incorporation of amino-acids into protein in the presence of polyribonucleotides consisting of only one type of base, they found that poly-U stimulates the incorporation of phenylalanine, while poly-A and poly-C do not stimulate the incorporation of any amino-acid. This means that phenylalanine corresponds to the triplet UUU, while the triplets CCC and AAA do not correspond to any amino-acid. It was also found that poly-UA stimulates the incorporation of Phe, Ileu, Tyr, Leu, Lys, and AspN; poly-UC, the incorporation of Phe, Pro, Leu, and Ser; poly-UG —Phe, Gly, Try, Leu, Val, and CySH; poly-UAC —Phe, Ser, Tyr, Leu, Ileu, Lys, AspN, Pro, Thr, and His; poly-UCG —Phe, Gly, Ala, Arg, Try, Val, CySH, and Leu; poly-UAG —Phe, Tyr, Gly, Met, Val, Asp, and AspN (Table III). Knowing that phenylalanine corresponds to the triplet UUU, we can determine the nucleotide composition of the triplets corresponding to particular amino-acids by varying the composition of the polyribonucleotide and comparing the incorporation of the different

Table III. The relative stimulation of the incorporation of amino-acids into protein by polyribonucleotides

Amino-acid	Ratio of incorporation into protein of phenylalanine to that of the given amino-acid upon addition of the given polyribonucleotides (averaged over several experiments)					
	UC 5 : 1	UA 5 : 1	UG 5 : 1	UAC 6 : 1 : 1	UGC 6 : 1 : 1	UAG 6 : 1 : 1
Ala.	—	—	—	—	31	—
Arg.	—	—	—	—	30	—
Asp.	—	—	—	—	—	41
AspN.	—	15	—	15	—	20
Val.	—	—	5	—	5	4
His.	—	—	—	29	—	—
Gly.	—	—	24	—	40	—
Glu.	—	—	—	—	—	64
Leu.	5	7	8	4	4	—
Ileu.	—	5	—	6	—	—
Lys.	—	32	—	46	—	—
Met.	—	—	—	—	—	23
Pro.	13	—	—	29	—	—
Ser.	4	—	—	64	—	—
Tyr.	—	4	—	4	—	5
Thr.	17	—	—	11	—	—
Try.	—	—	20	—	24	—
CySH.	—	—	5	—	4	—

amino-acids into protein with that of phenylalanine.

Since the base sequence in the artificial polyribonucleotide is random, the probability that a particular triplet contains three uracils, e.g., in the case of poly-UC, is obviously related to the probability that it contains two uracils and one cytosine by the ratio

$$\frac{C_U^3}{C_U^2 C_C} = \frac{C_U}{C_C},$$

where C_U and C_C are the relative contents of uracil and cytosine in the polymer.

The copolymer of uracil and cytosine, poly-UC, stimulates incorporation into protein not only of phenylalanine, but also of serine, leucine, threonine, and proline. For poly-UC having a ratio U/C = 5, the ratio of incorporation of phenylalanine to serine in the protein obtained is about five. Since phenylalanine is coded by the triplet UUU, thus we see that serine must correspond to a triplet consisting of two U's and one C. Here we still do not know the order of arrangement of U and C in the triplet. By this method, triplets corresponding to all 20 amino-acids have been found (Table IV). Not all the triplets corresponding to amino-acids are known yet, since the incorporation of the amino-acids was compared with that of phenylalanine, and correspondingly, only the triplets containing U were determined. However, it is already evident that meaningless triplets exist (AAA and CCC have been found thus far) and that the code is degenerate, since certain amino-acids correspond to several triplets. For example, the incorporation of leucine into the protein is stimulated both by poly-UC and by poly-UG, so that leucine must correspond to at least two triplets, one of which contains only U and C, and the other only U and G.

Table IV. The triplets corresponding to the amino-acids

Amino-acid	Triplets (order of nucleotides unknown)		
	1	2	3
Phe.		UUU	
Ala.		UGC	
Arg.		UGC	
Asp.	UAG		
AspN.	UAA, UAC		
Val.		UUG	
His.	UAC		
Gly.		UGG	
Glu.	UAG		UGC
GluN.	UGC*)		
Leu.	UUA	UUC, UUG	
Ileu.		UUA	
Lys.		UAA	
Met.	UAG		UGG
Pro.		UCC	
Ser.		UUC	UGC
Tyr.		UUA	
Thr.	UAC, UCC		
Try.		UGG	
CySH.		UUG	

1 – Triplets determined only by Ochoa; 3 – triplets determined only by Nirenberg and Matthaei; 2 – triplets determined both by Ochoa and by Nirenberg and Matthaei.

*) The triplet corresponding to glutamine, as identified by Ochoa from the substitution GluN → Val discovered by Tsugita[87] in TMV protein upon treatment of the RNA of this virus with nitrous acid.)

In order to determine the incorporation of the amino-acids into protein in the system, we must add all the amino-acids, with one of them bearing a radioactive label. The incorporation of this amino-acid is determined from the radioactivity of the acid-insoluble fraction of the protein obtained. Hence, the data obtained from analysis of the relative content of the different amino-acid residues in the polypeptides may in a number of cases be not completely reliable, and may require verification. The accuracy of these data may be improved by performing several parallel experiments using polynucleotides having various ratios of the bases they contain.

Certain supplementary information concerning the nucleotide composition of the triplets corresponding to the amino-acids can be obtained to verify that obtained by the biochemical method. This information is derived from data on amino-acid substitutions in the protein of the envelope of tobacco mosaic virus (TMV) in experiments like those of Schramm and Schuster described above.

Wittmann,[61] as well as other authors,[59,60] have found the following replacements of amino-acids in TMV protein upon treatment of the TMV RNA with nitrous acid:

Pro → Ser,
Pro → Leu,
Ser → Leu,
Ser → Phe,
Leu → Phe,
Thr → Ala,
Thr → Met,
Thr → Ileu,
Thr → Ser,
Ileu → Val,
Glu or GluN → Val,
Glu or GluN → Gly,
Asp or AspN → Ala,
Asp or AspN → Gly,
Asp or AspN → Ser.

Most of these transformations agree with the data on the nucleotide composition of the triplets corresponding to these amino-acids, as obtained by Ochoa and by Nirenberg and Matthaei. This, in particular, confirms the universality of the code. Only the replacement of threonine by methionine is discrepant, but we must bear in mind that the biochemical method has thus far determined only a fraction of all the triplets corresponding to the amino-acids.

Since the treatment of RNA with nitrous acid can bring about only the transformation of A into G and C into U, then, as Girer[83] has stated, all 64 triplets can be divided into eight groups of eight triplets each (the so-called octets) in such a way that the HNO_2-induced transformations of triplets into one another are confined within each octet (Fig. 6).

If the code is non-degenerate, so that each amino-acid is coded by only one triplet, the substitutions of the amino-acids will be divided into independent groups in concordance with the system of octets. An attempt by Wittmann[61] to relate the substitutions which he had observed to the system of octets showed that the code is degenerate, even before Ochoa had shown this, since he discovered transformations between amino-

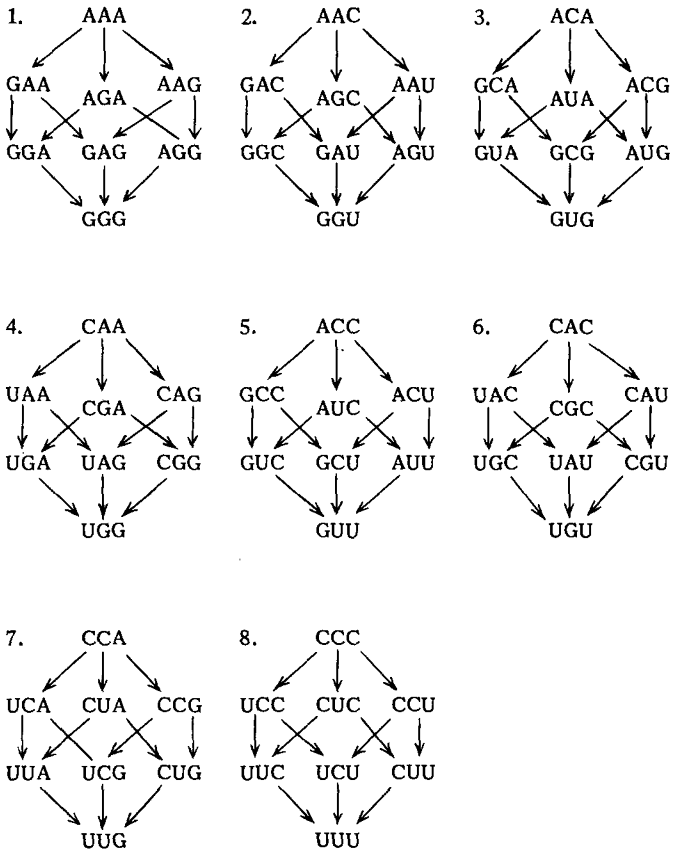

FIG. 6. The system of octets.[83]

acids which must belong to different octets. With the incomplete data which we have at present on the nucleotide compositions of the triplets and on amino-acid substitutions, we cannot yet make an unequivocal assignment of the amino-acids to the octets (with the exception of octet 8).

Thus, not only has the general nature of the biochemical code been determined at present, but also the nucleotide composition of a number of meaningful and meaningless triplets. However, the sequential order of the nucleotides in the triplets corresponding to particular amino-acids has not yet been determined. In principle, this order cannot be determined by use of the artificial polyribonucleotides obtained by Ochoa's method, since in the process of their biosynthesis, the particular nucleotides are arranged quite at random along the polymer chain. The situation would be quite different if we could obtain by chemical synthesis polyribonucleotides having an assigned arrangement of bases along the chain. However, this problem has not been solved yet.

It seems possible to determine the order of the nucleotides in the triplets by studying the replacement of a region of the polypeptide chain of the protein in double acridine mutants, as obtained by experiments like those of Crick and his associates. Indeed, when a phage contains both a direct mutation and a suppressor for it, as has been stated above, the region of the DNA molecule between the mutation and its suppressor exhibits a shift in the readout of the genetic information. This is manifested in the fact that an amino-acid in the altered region of the polypeptide chain of the protein is determined by a triplet obtained by the overlap of two triplets corresponding to two successive amino-acids in the protein of the original wild-type phage. If we know the amino-acid sequence in the protein of the original type and in the protein of the double acridine mutant, as well as the nucleotide composition of all the triplets corresponding to the individual amino-acids, obviously, we may obtain a set of data permitting us to determine the order of the nucleotides in the triplets. However, in order to do this, we must know just which protein corresponds to the particular region in the genetic material containing the suppressors of the acridine mutations. Furthermore, we must be able to isolate and purify this protein from even small quantities of other proteins as impurities. Only after this can we determine the amino-acid sequence in either the wild-type protein or in the proteins of the double mutants.

Perhaps the difficulty associated with the lack of a method of synthesizing polynucleotides containing a known base order may be avoided if, as K. S. Mikhaĭlov has suggested, we attach a known group of nucleotides, or even a single nucleotide to the end of a polyribonucleotide like poly-U, and determine the terminal amino-acids in the polypeptide synthesized by this polynucleotide in the Nirenberg-Matthaei system. Unfortunately, however, we have no a priori knowledge whether the terminal groups of polyribonucleotides participate in protein synthesis in such a system.

The only biosynthesized polynucleotide having a known base sequence is polydeoxy-AT. This copolymer, which is synthesized by DNA-polymerase, has been isolated by Kornberg, and has a regularly alternating sequence of A and T. The copolymer polyribo-AU synthesized on this template using RNA-polymerase also possesses a regularly alternating sequence of A and U. In the Nirenberg-Matthaei system, such a polyribonucleotide must stimulate the incorporation of amino-acids coded by triplets of quite definite nucleotide sequence, AUA and UAU, provided that these triplets are not meaningless. The triplet UAU corresponds either to tyrosine, or to leucine, or to isoleucine. Among the three triplets consisting of U+2A, at least two correspond to amino-acids, asparagine and lysine. This permits us to hope for a positive result from this experiment, which might clear up the nucleotide order in two triplets.

Thus we can envision ways of determining the nucleotide order in the triplets, and we have every ground for expecting that very soon the biochemical code relating the sequence of amino-acid residues in a protein with the base sequence in the region of a DNA molecule determining the structure of this protein will be completely solved.

Besides the determination of the type of the biochemical code and the exact meanings of the individual triplets which occur in it, considerable advances have been made recently in determining the complex mechanism controlling the reading of this code. This mechanism is a component part of the system controlling protein synthesis in the cell, and seems, above all, to affect the synthesis of mRNA. Thus, e.g., Khesin and Shemyakin[84] have recently found that when a bacterium is infected by a phage, the mRNA which is synthesized in the bacterium on the introduced phage DNA differs at different stages of development of the phage. In the first minutes after the bacterium has been infected by the phage, the mRNA is synthesized on certain regions of the DNA, and later on others.

The subtle genetic experiments conducted in the last few years by Jacob and Monod have shown that there are cases in which, in addition to the structural gene, which serves as a template for mRNA synthesis and determines the amino-acid sequence in the corresponding protein, there is also an operator gene arranged alongside the structural gene and apparently directly exerting control by synthesizing mRNA. Besides, there is also a regulator gene which can be located in a different part of the genome. This gene exerts control by synthesizing some special substance (possibly also one of the fractions of RNA) called the repressor, affecting the function of the operator gene.[85]

An extremely interesting and thus far unsolved problem is that of the "period" dividing one meaningful "sentence" from another in the DNA molecule corresponding to the structural gene. This problem is closely related to the determination of the mechanism of action of the enzyme which copies the information from the DNA onto the mRNA (RNA polymerase). At present the mechanisms of action of enzymes bringing about synthesis on a template have not been studied at all.

Three processes of synthesis on a template are known: reduplication of DNA, synthesis of mRNA on the DNA, and synthesis of protein on the mRNA. All these processes are characterized by the fact that the template is an individual strand of nucleic acid, and the synthesis proceeds sequentially in a particular direction fixed by the template. Nothing is known about the action of this type of enzyme, and we can only advance some hypotheses, which do not pretend at all to be true. Possibly a single molecule of the enzyme attaches itself to the template and moves along it. This would permit us to explain, in particular, why synthesis of an entire new molecule always takes place, and we do not observe detachment of only partially synthesized molecules from the template. Possibly in the synthesis of mRNA on DNA, the enzyme RNA polymerase is attached to a particular "period" separating adjacent independently-functioning regions of the DNA (*cistrons*), and follows along the DNA strand until it meets the next period. These periods, whose existence has been demonstrated by Crick, are possibly specific sequences of several nucleotides formed of the usual four bases, or they may contain certain bases of rare occurrence in the molecule. In the synthesis of mRNA on the DNA, the synthesis possibly takes place of both strands of mRNA, complementary to both strands of the DNA (it is not known whether synthesis of the two strands of mRNA takes place simultaneously or successively). However, it seems most probable that only one of the complementary strands of mRNA can act as the template for protein synthesis, and hence, only this mRNA strand can be incorporated into the ribosomes. The fate of the second strand is not yet clear. Perhaps these complementary mRNA strands are distinguished by their end groups, as determined by the nature of the "periods" in the DNA.

Berg[86] has shown recently that, at least in a special system consisting of ribosomes isolated from the cell, supernatant liquid, and a set of enzymes, the only mRNA which can participate in protein synthesis is that synthesized on two-stranded DNA molecules; the mRNA synthesized on denatured DNA does not increase the incorporation of amino-acids into the ribosome proteins. On the basis of these data, we may suppose that the period in the DNA functions normally only in the two-stranded state, although mRNA synthesis also takes place separately on each strand.

At present broad investigations are being conducted throughout the world on the details of the systems regulating protein synthesis, and we may expect that new major advances will occur in this field, even in the next few years.

[1] J. D. Watson and F. H. C. Crick, Nature **171**, 737 (1953).

[2] F. H. C. Crick and J. D. Watson, Proc. Roy. Soc. (London) **A223**, 80 (1954).

[3] Josse, Kaiser, and Kornberg, J. Biol. Chem. **236**, 864 (1961).

[4] Langridge, Wilson, Hooper, Wilkins, and Hamilton, J. Mol. Biol. **2**, 19 (1960).

[5] Chargaff, Crampton, and Lipshitz, Nature **172**, 289 (1953).

[6] N. Sueoka, J. Mol. Biol. **3**, 31 (1961).

[7] E. Chargaff, in collected volume Biological Structure and Function, Vol. 1, p. 67, Academic Press, New York (1961).

[8] C. Levinthal and P. F. Davison, J. Mol. Biol. **3**, 674 (1961).

[9] J. Cairns, J. Mol. Biol. **3**, 756 (1961).

[10] J. K. Setlow and R. B. Setlow, Proc. Natl. Acad. Sci. U.S.A. **46**, 791 (1960).

[11] R. L. Sinsheimer, J. Mol. Biol. **1**, 43 (1959).

[12] J. D. Watson and F. H. C. Crick, Nature **171**, 964 (1953).

[13] J. D. Watson and F. H. C. Crick, Cold Spring Harbor Symposia Quant. Biol. **18**, 123 (1953).

[14] M. Meselson and F. W. Stahl, Proc. Natl. Acad. Sci. U.S.A. **44**, 671 (1958).

[15] J. Marmur and D. Lane, Proc. Natl. Acad. Sci. U.S.A. **46**, 453 (1960).

[16] Doty, Marmur, Eigner, and Schildkraut, Proc. Natl. Acad. Sci. U.S.A. **46**, 461 (1960).

[17] J. Marmur and P. Doty, J. Mol. Biol. **3**, 585 (1961).

[18] Schildkraut, Marmur, and Doty, J. Mol. Biol. **3**, 595 (1961).

[19] Lehman, Bessman, Simms, and Kornberg, J. Biol. Chem. **233**, 163 (1958).

[20] Bessman, Lehman, Simms, and Kornberg, ibid. **233**, 171 (1958).

[21] Adler, Lehman, Bessman, Simms, and Kornberg, Proc. Natl. Acad. Sci. U.S.A. **44**, 641 (1958).

[22] A. Kornberg, Science **131**, 1503 (1960).

[23] Doty, Boedtker, Fresco, Haselkorn, and Litt, Proc. Natl. Acad. Sci. U.S.A. **45**, 482 (1959).

[24] A. S. Spirin, J. Mol. Biol. **2**, 436 (1960).

[25] F. H. Crick, Symposia Soc. Exptl. Biol. **12**, 138 (1958), Acad. Press, New York.

[26] H. K. King, Sci. Progr. **49**, 703 (1961).

[27] B. D. Hall and P. Doty, J. Mol. Biol. **1**, 111 (1959).

[28] C. G. Kurland, J. Mol. Biol. **2**, 83 (1960).

[29] Tissières, Watson, Schlessinger, and Hollingworth, J. Mol. Biol. **1**, 221 (1959).

[30] P. C. Zamecnik and E. B. Keller, J. Biol. Chem. **209**, 337 (1954).

[31] M. B. Hoagland, Biochim. Biophys. Acta **16**, 288 (1955).

[32] Hoagland, Keller, and Zamecnik, J. Biol. Chem. **218**, 345 (1956).

[33] Hecht, Stephenson, and Zamecnik, Proc. Natl. Acad. Sci. U.S.A. **45**, 505 (1959).

[34] Hecht, Zamecnik, Stephenson, and Scott, J. Biol. Chem. **233**, 954 (1958).

[35] Bishop, Leahy, and Schweet, Proc. Natl. Acad. Sci. U.S.A. **46**, 1030 (1960).

[36] H. M. Dintzis, Proc. Natl. Acad. Sci. U.S.A. **47**, 247 (1961).

[37] H. Schuster and G. Schramm, Z. Naturforschung **13b**, 697 (1958).

[38] Tsugita, Gish, Young, Fraenkel-Conrat, Knight, and Stanley, Proc. Natl. Acad. Sci. U.S.A. **46**, 1463 (1960).

[39] E. Volkin and L. Astrachan, Virology **2**, 149, 433 (1956).

[40] E. Volkin and L. Astrachan, The Chemical Basis of Heredity, ed. W. D. McElroy and B. Glass, p. 686, Johns Hopkins Press, Baltimore (1957); Russ. Transl. Khimicheskie osnovy nasledstvennosti, p. 556, M., IL (1960).

[41] Nomura, Hall, and Spiegelman, J. Mol. Biol. **2**, 306 (1960).

[42] M. Yčas and W. S. Vincent, Proc. Natl. Acad. Sci. U.S.A. **46**, 804 (1960).

[43] Gros, Hiatt, Gilbert, Kurland, Risebrough, and Watson, Nature **190**, 581 (1961).

[44] Risebrough, Tissières, and Watson, Proc. Natl. Acad. Sci. U.S.A. **48**, 430 (1962).

[45] Brenner, Jacob, and Meselson, Nature **190**, 576 (1961).

[46] B. D. Hall and S. Spiegelman, Proc. Natl. Acad. Sci. U.S.A. **47**, 137 (1961).

[47] Spiegelman, Hall, and Storck, ibid. **47**, 1135 (1961).

[48] H. M. Schulman and D. M. Bonner, ibid. **48**, 53 (1962).

[49] S. B. Weiss, Proc. Natl. Acad. Sci. U.S.A. **46**, 1020 (1960).

[50] Hurwitz, Bresler, and Diringer, Biochem. Biophys. Res. Communs. **3**, 15 (1960).

[51] Ochoa, Burma, Kröger, and Weill, Proc. Natl. Acad. Sci. U.S.A. **47**, 670 (1961).

[52] Furth, Hurwitz, and Goldmann, Biochem. Biiphys. Res. Communs. **4**, 362 (1961).

[53] M. Chamberlin and P. Berg, Proc. Natl. Acad. Sci. U.S.A. **48**, 81 (1962).

[54] G. Gamow, Biol. Med. Danske Vid. Selskab. **22**, No. 3 (1954).

[55] G. Gamow, Nature **173**, 318 (1954).

[56] Gamow, Rich, and Yčas, Advances in Biol. Med. Phys. **4**, 23 (1956), Academic Press, New York.

[57] M. Yčas, in Symposium on Information Theory in Biology, Gatlinburg, Tenn., Oct. 29-31, 1956, ed. H. P. Yockey, Pergamon Press, New York (1958); Russ. Transl. in collected volume Teoriya informatsii v biologii (Information Theory in Biology), p. 72, M., IL (1960).

[58] S. Brenner, Proc. Natl. Acad. Sci. U.S.A. **43**, 687 (1957).

[59] A. Tsugita and H. Fraenkel-Conrat, Proc. Natl. Acad. Sci. U.S.A. **46**, 636 (1960).

[60] A. Tsugita and H. Fraenkel-Conrat, J. Mol. Biol. **4**, 73 (1962).

[61] H. G. Wittmann, Naturwiss. **48**, 730 (1961).

[62] N. Sueoka, Proc. Natl. Acad. Sci. U.S.A. **47**, 1141 (1961).

[63] G. Gamow, Biol. Med. Danske Vid. Selskab **22**, No. 8 (1954).

[64] G. Gamow and M. Yčas, Proc. Natl. Acad. Sci. U.S.A. **41**, 1011 (1955).

[65] A. L. Dounce, Enzymologia **15**, 251 (1952).

[66] M. Yčas, Naturwiss. **43**, 197 (1956).

[67] Crick, Griffith, and Orgel, Proc. Natl. Acad. Sci. U.S.A. **43**, 416 (1957).

[68] V. V. Chavchanidze, Biofizika **3**, 391 (1958); Engl. Transl., Biophysics **3**, 377 (1958).

[69] Golomb, Welth, and Delbruck, Biol. Med. Danske Vid. Selskab **23**, No. 9 (1958).

[70] M. W. Nirenberg and J. H. Matthaei, 5th Congress of the International Union of Biochemistry, Moscow (1961).

[71] M. W. Nirenberg and J. H. Matthaei, Proc. Natl. Acad. Sci. U.S.A. **47**, 1588 (1961).

[72] Nirenberg, Matthaei, and Jones, ibid. **48**, 104 (1962).

[73] Crick, Barnett, Brenner, and Watts-Tobin, Nature **192**, 1227 (1961).

[74] S. Benzer, see [40], p. 70; Russ. Transl., p. 56.

[75] Brenner, Barnett, Crick, and Orgel, J. Mol. Biol. **3**, 121 (1961).

[76] J. L. Jinks, Heredity **16**, 153, 241 (1961).

[77] S. Ochoa and L. A. Heppel, see [40], p. 615; Russ. Transl., p. 500.

[78] Lengyel, Speyer, and Ochoa, Proc. Natl. Acad. Sci. U.S.A. **47**, 1936 (1961).

[79] Speyer, Lengyel, Basilio, and Ochoa, ibid. **48**, 63 (1962).

[80] Lengyel, Speyer, Basilio, and Ochoa, ibid. **48**, 282 (1962).

[81] Speyer, Lengyel, Basilio, and Ochoa, ibid. **48**, 441 (1962).

[82] Martin, Matthaei, Jones, and Nirenberg, Biochem. Biophys. Res. Communs. **6**, 410 (1962).

[83] A. Girer, 5th Congress of the International Union of Biochemistry, Symposium III, Moscow (1961).

[84] M. F. Shemyakin and R. B. Khesin, DAN SSSR (1962) (in press).

[85] F. Jacob and J. Monod, J. Mol. Biol. **3**, 318 (1961).

[86] W. B. Wood and P. Berg, Proc. Natl. Acad. Sci. U.S.A. **48**, 94 (1962).

[87] A. Tsugita, Protein, Nucleic Acid, Enzyme (Tokyo) **6**, 385,(1961).

Translated by M. V. King