UNPUBLISHED PRELIMINARY DATA

PRESENT STATUS OF THE AMINO ACID CODE

Thomas H. Jukes
Space Sciences Laboratory
University of California
Berkeley, California

Presented at the annual meeting of the American Dietetic Association, Portland, Oregon, July 29, 1964

*Supported by a grant from the National Aeronautics and Space Administration (NSG 479)

The fact that you have asked me to speak on the amino acid code at your annual meeting indicates the great interest that this subject has aroused among people whose work is centered in the biological sciences. It is appropriate that this should be so because the code is believed to be related directly or indirectly to the synthesis of all the proteins that are found in every form of life. It is proteins that are responsible for the differences between various living species, and for the biological synthesis of all non-protein molecules, since these are produced by enzymes, which are themselves proteins.

All of us as students of nutrition are familiar with the fact that proteins are composed of amino acids, and that the quantities of amino acids vary in the different proteins. In addition to this general principle, it is known that each protein contains a definite sequence formed by linking together various numbers of the 20 amino acids which are shown in Table 1. The order in which they are arranged is controlled by the sequence of 4 substances in molecules of deoxyribonucleic acid (DNA). These 4 substances are the bases adenine, cytosine, guanine and thymine (A,C,G,T). This conclusion has far-reaching implications. One of these is that the difference between two species, such as a human being and a pine tree, is due

solely to the number and sequence of the 4 bases, A,C,G and T in DNA. This follows from the fact that all animals and higher plants grow from fertilized ova in which the only essential difference is the base composition of their DNA. We can go even further in our analogy; the same rule applies to yeasts, bacteria and even to such small particles as the smallpox virus. We are confronted with a unifying principle in biology and the task before us is to translate it through its various complexities.

The molecule of DNA is fully equal to these impressive obligations. Figure 1 is a diagram showing a portion of a DNA molecule. Its structure resembles that of a ladder. The two uprights of the ladder are chains made of phosphate groups joined by molecules of deoxyribose, a 5-carbon-atom sugar. The chains run in opposite directions, as shown in figure 1. The rungs of the ladder are pairs of bases joined by hydrogen bonds, such that adenine is always paired with thymine and cytosine with guanine. These rungs, or units, may be placed in a virtually infinite number of different sequences. This is illustrated in Table 2. The size of such numbers is illustrated by the old fable about the inventor of the game of chess, who was said to have been asked by his King to name a reward for the achievement. The man replied that all he would

ask was one grain of rice for the first square of the chessboard, two for the second, four for the third, doubling each time until the sixty-fourth square had been reached. The King immediately granted this simple request, only to find that the requirement was unmanageable; indeed, the sixty-fourth square alone needed 2^{64} grains of rice, which would weigh more than 1600 billion tons. Similarly, a sequence of 64 letters, using A,C,G and T in all possible combinations, can be written in 4^{63} different ways. This number is even more impressive than the example from the chessboard; it is about 2 billion billion billion billion. Such a sequence, in terms of DNA, would represent only about one-seventh of the length of a typical gene. Obviously these vast numbers of different possibilities cannot exist in nature. Only a handful of combinations can survive the process of natural selection that takes place during evolution. The surviving molecules are those that carry the information leading to the production of viable combinations of proteins.

It is believed that the final step in the synthesis of proteins is actually directed by a second form of nucleic acid, ribonucleic acid (RNA), rather than by DNA. We shall discuss this later but one reason for this conclusion is that proteins are synthesized in the cytoplasm of animal cells rather than in the nucleus, which contains almost all the DNA. A molecule formed by combining a single base with ribose is called a nucleoside. RNA is made of nucleosides linked together by phosphates.

The amino acid code, sometimes termed the genetic code, is the key by which the sequence of bases in DNA is translated into amino acids during the synthesis of proteins. The amino acid code was solely a matter of theory until 1961 when a crucial experiment with a completely unpredictable result was carried out by Nirenberg and Matthaei at the National Institutes of Health. These two scientists were investigating the synthesis of protein in extracts prepared from cells of the common intestinal organism Escherichia coli. Nirenberg and Matthaei added an artificial form of RNA to the extract. The artificial form contained only one base, uracil, in long strands of a compound called polyuridylic acid. The experiment showed by radioactive tracer techniques that a new protein-like substance was formed by the E. coli enzymes in the extract. This substance contained only one amino acid, phenylalanine. The code for phenylalanine therefore consisted only of uracil. (1)

This famous experiment bridged the last remaining gap that had separated theoretical genetics and test-tube biochemistry. It now became experimentally possible to search for codes for the other 20 amino acids (Table 1), that are concerned in the synthesis of proteins.

It had previously been concluded that the code did not consist of either 1 or 2 bases per amino acid, simply because there are 20 different amino acids and only 4 different bases. There are only

16 different ways of arranging A,C,G and T in pairs. For this and other reasons it is thought that a triplet of 3 consecutive bases is needed to code for each amino acid. It is known that the sequences of bases in either of the two strands of a DNA molecule are unrestricted with respect to the order in which they occur: apparently any one of the four bases can be next to any of the other four, although of course each base must be paired with the corresponding complementary base in the adjacent strand as shown in Fig. 1. The same lack of restriction is true of the amino acid sequences in the polypeptide chains of proteins. Any of the 20 amino acids can occur next to any other. Moreover, the sequences in DNA are subject to mutational changes in which one base replaces another, or bases are added to or deleted from the DNA. Such rearrangements plus the possibility of lengthening of DNA molecules can be numerous enough to account for all the evolution of living forms since the first appearance of life on earth.

Let us now turn to the procedure by which the information entrained in the ultramicroscopic molecule of DNA is translated into proteins. The DNA molecule contains 2 strands which are held together by hydrogen bonds that bridge the gaps between adenine and thymine or between guanine and cytosine as shown in Figure 1. This is called the "Watson-Crick" pairing mechanism from the names of its discoverers. (2) The existence of this copying mechanism was ac-

tually predicted by Watson and Crick before it was shown to exist in living systems. It means that when a new strand of DNA is formed, the copying mechanism is such that A pairs with T and G pairs with C as shown in Figure 2. This is the procedure by which hereditary characteristics are transmitted from parents to children. A second copying mechanism exists by which these characteristics are expressed in the visible attributes of an individual that are hereditary in origin. The first step in this second procedure is a modified copying or transcribing process in which the double strand of DNA is used as a template by an enzyme that assembles nucleoside triphosphates into molecules of RNA. Again, the sequence of bases is repeated in a complementary manner but this time A pairs with a fifth base, uracil (U) instead of thymine. Apparently in living cells only one strand of the DNA is "transcribed" by the enzyme that makes RNA. We can call this strand of DNA the "minus" strand as shown in Figure 3, which diagrammatically illustrates the "transcribing" process.

The RNA molecules are of several types and are usually divided into three classes. The first class of molecules is called messenger RNA. These are fairly long molecules and each one carries a specific message which will decide the number and order of amino acids in a protein. Some of them carry the messages for several proteins which

are made consecutively. The second type of RNA molecules is called s-RNA or transfer RNA. These are quite short; each contains about 70 bases. Each of these has the function of combining with a specific amino acid to be placed in the right spot in a new protein. The third type of RNA is ribosomal RNA, which has a structural function. It combines with ribosomal protein to form particles called ribosomes. These become joined together with strands of messenger RNA as polysomes, which resemble ultramicroscopic strings of beads, made of ribosomes strung together by RNA.

The s-RNA molecules bring the attached amino acids to the polysomes and arrange the amino acids in order along the strands of messenger RNA. Peptide linkages are then formed, one by one, and a new protein results. Let us now turn to the coding procedure by which this takes place. Most of our knowledge is due to experiments with synthetic RNA carried out with extracts of E. coli as previously outlined. The majority of the work has been at Dr. Nirenberg's laboratory at the National Institutes of Health and at Professor Ochoa's laboratory at New York University. Various combinations of A,C,G and U were used in preparing the synthetic RNA molecules that are used in experiments to explore the code. These molecules are made by incubating a mixture of ribonucleoside diphosphates with a specific enzyme, polynucleotide phosphorylase. An important property of this enzyme is that it arranges the bases into polynucleotide strands containing

random sequences depending upon the proportion of each base. For example, if the enzyme were furnished with a mixture of 5 parts of A and 1 part of C it would make strands containing on the average 25 sequences of AAA, 5 of AAC, 5 of ACA, 5 of CAA, and one each of ACC, CAC and CCA. This principle is illustrated in Table 3 and is important in making deductions regarding the code. The proportion of triplets within the strands of a polynucleotide is reflected in the proportion of amino acids in polypeptides that are obtained in the cell-free system. It is upon this concept that most of the present knowledge of the amino acid code is based (3,4,5). In Table 4 are summarized all the proposed codes that have been discovered by this experimental approach.

The list is not yet complete. There are 64 possible permutations that can be written using the 4 letters A,C,G and U, and a total of only about 48 have been assigned so far. The missing assignments are indicated as blanks in Table 4, for example, 1A, 1C, 1G can be written as ACG, AGC, CAG, CGA, GAC, GCA and only 3 amino acids have so far been found to be coded by 1A, 1C, 1G. On the other hand, six amino acids are reported to be coded by 1A, 1C, 1U, as shown in Table 4.

It is believed that the delay in finding more about the other 16 "triplets" is due in part to the fact that these 16 all contain G which tends to form cross-linkages in the long molecules. This results in the strands becoming tangled so that the message is lost.

It is also probable that a few of the triplets do not code for amino acids, but instead produce spaces between protein molecules.

The experiments with synthetic polynucleotides in test-tubes were made under conditions very different from the delicate and intricate processes that take place in living cells. Could the code found in the laboratory experiments have any relation to the system used by live organisms? A number of experimental observations indicate that there is at least a partial relationship.

This second group of experiments was concerned with the chemistry of mutations. One of the origins of this approach goes back to studies by Sanger and his collaborators on the chemistry of insulin (7). This investigation showed that the amino acids in a specific protein molecule did not vary in kind or number and that they were always arranged in the same linear sequence. However, when two insulins from different species were compared, such as beef and pork insulins, there was found to be a difference in two or three of the amino acids and it was always the same amino acids that varied in the same way. The characteristic differences between a cow and a pig therefore included changes in the amino acid sequence of insulin and, as was later shown, of other proteins as well. Such changes were also discovered in mutations within a species and it is most interesting that the first discovery was made with human beings who are suffering from a hereditary disease, termed sickle-cell anemia.

The disease was shown to be caused by a single amino acid change in the hemoglobin molecule (8). This molecule contains two identical "alpha" polypeptide chains and two identical "beta" polypeptide chains, each containing about 146 amino acids. Normal people have a glutamic acid group in the sixth link of the amino acid chain in beta hemoglobin. Sickle-cell anemia patients have valine in this position. This single amino acid change in a sequence of 146 amino acids has such a profound effect on the properties of hemoglobin that a fatal anemia can result. It was correctly predicted by Ingram that the change of a single nucleic acid base would suffice to bring about the substitution of glutamic acid by valine (9). Table 4 shows that this prediction is realized; the code for valine is 2U, 1G and a code for glutamic acid is 1A, 1U, 1G.

Here, therefore, is an explanation of mutations. They can be due to changes in protein molecules. These changes are a reflection of alterations in the number and kind of bases at specific locations in the molecule of DNA. Unacceptable changes are quickly eliminated by the process of natural selection. At rare intervals a change takes place which improves the characteristics of its possessor. These changes will survive and by such procedures we presume that in some cases new species have evolved.

By painstaking analysis of the proteins in which mutational changes have taken place in various species, about 38 different examples have been discovered where each amino acid changes to another

in the manner mentioned above for sickle-cell hemoglobin. This is a very interesting list for various reasons; the first being that the list includes examples taken from human hemoglobin, from bacterial cells, and from tobacco mosaic virus which is perhaps the most primitive form of life that has been detected. Second, nearly all the examples correspond to single-base changes in the proposed coding triplets in Table 4. It is indeed striking that there should be so close a similarity between two sets of results obtained quite independently. This encourages one to think that the test tube experiments with synthetic nucleic acids have produced information about the code that correspond fairly closely to what happens in living organisms. The next question, and a most important one, is to find out the order in which the bases are placed in the individual triplet coding units. There is only one way to write a triplet containing three A's but there are three ways of writing 2A and 1C and six ways of writing 1A, 1C and 1U. Furthermore, 3 different amino acids are coded by 2 A's and 1C and 6 different amino acids are coded by 1A, 1C and 1U. Therefore, the 3 possible ways of writing 2 A's and 1C which are AAC, ACA, and CAA must be assigned to asparagine, threonine and glutamine.

The single-amino-acid mutations are listed in Table 5. These have been found in proteins from a wide range of species; some in

human hemoglobin, some in bacteria, and some in the protein of the tobacco mosaic virus, which is representative of the smallest and simplest forms of parasitic life. Table 5 includes the work of many different laboratories, as reviewed elsewhere (5). The changes in tobacco mosaic virus, except those indicated as spontaneous, were produced by treatment with nitrous acid. It is striking that in several cases the same exchange between two amino acids has occurred both in human beings and the tobacco mosaic virus, hinting that the same coding mechanism occurs throughout living organisms.

In terms of the amino acid code, the changes in Table 5 should represent a change of only one of the three bases in a coding triplet, since the chances are overwhelmingly against two such changes taking place almost simultaneously. In 36 out of the 38 examples, the amino acid change does indeed correspond to a single-base change in the triplets in Table 4. Furthermore, as explained elsewhere (5), the bases in each triplet may be placed in an order which corresponds to the changes that take place in the mutations; for example if valine is written as UUG, one of the isoleucine codes should be written as UUA to account for the change A to G in the mutation isoleucine to valine. Space does not permit a discussion of this question, nor of the theories regarding the mechanism by which sRNA brings the amino acids to their correct positions on the messenger RNA in the polysomes.

The findings reviewed in this short article are enough to show that some insight is possible into the important biological problem of the genetic code. It is evident that only a beginning has been made. Certain nutritional problems are illuminated by the new findings; we can now see why growth stops so rapidly if an essential amino acid is withheld from the diet. Such a lack will immediately stop the growing cells from completing the pre-ordered sequences of amino acids in the chains of proteins. We can also understand why a protein such as zein can never contain tryptophan - it means that the gene in corn that makes zein simply does not furnish any coding triplets for tryptophan.

In the field of space science, there is an interest in problems in biology, such as the origin and evolution of life from non-living matter, and the possible existence of life elsewhere than on the earth. These problems confront us with the most fundamental biological questions. When did the amino acid code first appear? Can any organisms evolve without the code? For life to appear in other worlds, are nucleic acid and protein necessary, or can some other chemical systems suffice?

Finally, we realize that the genetic code illustrates one of the great principles of nature. This is the principle that complex entities are built by repetition and combination of a few small fundamental units. The unraveling of such complexities confronts us as a monumental task, but the biological sciences are to-day drawn togeth-

er in closer union by the perception that the maintenance of life is centered around the properties of nucleic acids and proteins.

References

(1) Nirenberg, M. W. and Matthaei, J. H. Proc. Natl. Acad. Sci. 47, 1588 (1961).

(2) Watson, J. D. and Crick, F. H. C. Nature 171, 737 (1953).

(3) Ochoa, S. Experientia 20, 57 (1964).

(4) Nirenberg, M. W., Jones, O. W., Leder, P., Clark, B. F. C. Sly, W. S., and Pestka, S. Cold Spring Harbor Symposia on Quantitative Biology, 28, 549 (1963).

(5) Jukes, T. H. American Scientist 51, 227 (1963).

(6) Speyer, J. F., Lengyel, P., Basilio, C., Wahba, A. J., Gardner, R. S. and Ochoa, S. Cold Spring Harbor Symposia on Quantitative Biology, 28, 559 (1963).

(7) Sanger, F. and Tuppy, H. Biochem. J. 49, 481 (1951).

(8) Ingram, V. M. Nature 180, 326 (1957).

(9) Ingram, V. M. Scientific American 204, 357 (1958).

Table 1. The amino acids that take part in protein synthesis and their abbreviations

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

Table 2.

Possible Numbers of Different Permutations of A,C,G and T

One letter $(4^1) = 4$ (A,C,G,T)

Two letters $(4^2) = 16$ (AA,AC,CA,AG,GA,AT,TA,CC,GC,CG,...etc)

Three letters $(4^3) = 64$ (AAA,AAC,ACA,CAA...etc)

Four letters $(4^4) = 256$ (AAAA,AAAC,AACA...etc)

10 letters $(4^{10}) = 1,048,000$

400 letters (4^{400}) (one gene)=1,209 million billion billion

$4^{3,000,000,000}$ letters-The practical equivalent of infinity.

(Mammalian DNA has 3,000,000,000 bases)

Table 3. Calculated triplet frequencies and observed amino acid incorporation with synthetic ribonucleotides containing 5 parts of A and 1 part of C (From Speyer and co-workers (5)).

Calculated Triplet Frequency	Relative Amino-Acid Incorporation*		Coding Assignments
AAA - 100	Lysine	100	AAA
CAA - 20	Threonine	26.5	ACA and CCA
ACA - 20	Asparagine	24.2	CAA
AAC - 20	Glutamine	23.7	AAC
CCA - 4	Proline	7.2	CAC and CCC
CAC - 4	Histidine	6.5	ACC
ACC - 4			
CCC - 0.8			

*Calculated on basis of lysine = 100

Table 4. The amino acid code: RNA triplets that have been found to correspond to the incorporation of amino acids into polypeptides. The abbreviations for the amino acids are in Table 1. Unassigned possibilities are indicated by -.

3A Lys	1C,2G Gly - -
3C Pro	1U,2G,Try,gly -
3G -	1C,2U Ser,phe,leu
3U Phe	1G,2U Cys,leu,val
2A,1C GlN,thr,asN	1G,2C Arg,ala,thr
2A,1G Glu,arg -	1U,2C Ser,pro,leu
2A,1U Ilu,lys,asN	1A,1C,1G Ala,asp,ser - - -
1A,2C His,pro,thr	1A,1C,1U Tyr,his,ilu,asN,glN,thr
1A,2G Gly,glN -	1A,1G,1U Glu,asp,met - - -
1A,2U Tyr,leu,ilu	1C,1G,1U Ala,arg - - - -

Table 5. Proteins and the Amino Acid Code; Single Amino Acid Changes in Mutations Compared with Single Base Changes in Messenger RNA in Terms of Proposed Coding Triplets.

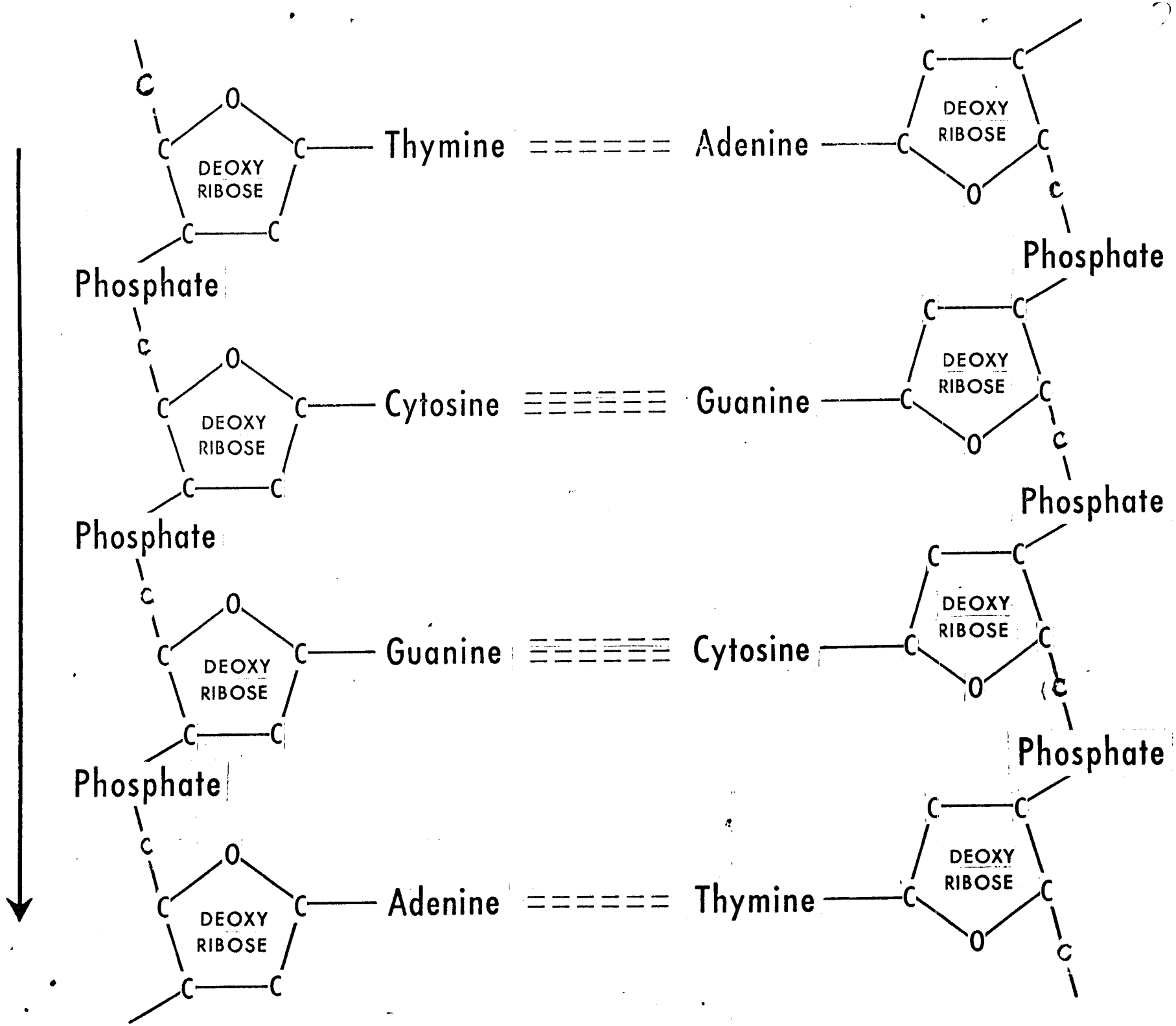
Protein	Amino Acid Change	Corresponding Base Change	Protein	Amino Acid Change	Corresponding Base Change
Hb,TMV**	Lys/asN	A/C,A/U	TMV	Thr/ser	A/G,A/C
Hb	Lys/asp	A/G	TMV	Thr/met	C/G
Hb	Lys/glu	A/G	TMV,TS	Thr/ilu	C/U,C/A
TMV	AsN/ser	A/U	TMV	Pro/leu	C/U
TMV**	AsN/arg	C/G,U/G	TMV	Pro/ser	C/U
TMV**	AsN/asp	C/G	L,AP	Ala/val	C/U
Hb	His/tyr	C/U	TMV,TS	Ser/leu	C/U
Hb	His/arg	A/G	TMV	Ser/phe	C/U
L	His/glN	C/A,U/A	TMV	Asp/ala	(G*A/C*G)
Hb	GlN/arg	G/A	TMV,Hb,L,TS	Asp/gly	A/G
TS	Glu/ala	A/C	TS	Gly/val	G/U
Hb	Glu/glN	C/G	TS	Gly/cys	G/U
Hb,TMV,TS	Glu/gly	A/G	TMV,TS	Arg/gly	C/G,A/G
Hb,TS	Glu/val	A/U	TS	Arg/ser	C/G
TS	Tyr/cys	A/G	TS	Arg/leu	G/U
TMV	Tyr/phe	A/U	TS	Arg/thr	?
Cyt c	Met/leu	A/U	TMV	Ilu/val	A/G
TMV	Leu/phe	C/U	TMV	Ilu/met	U/G

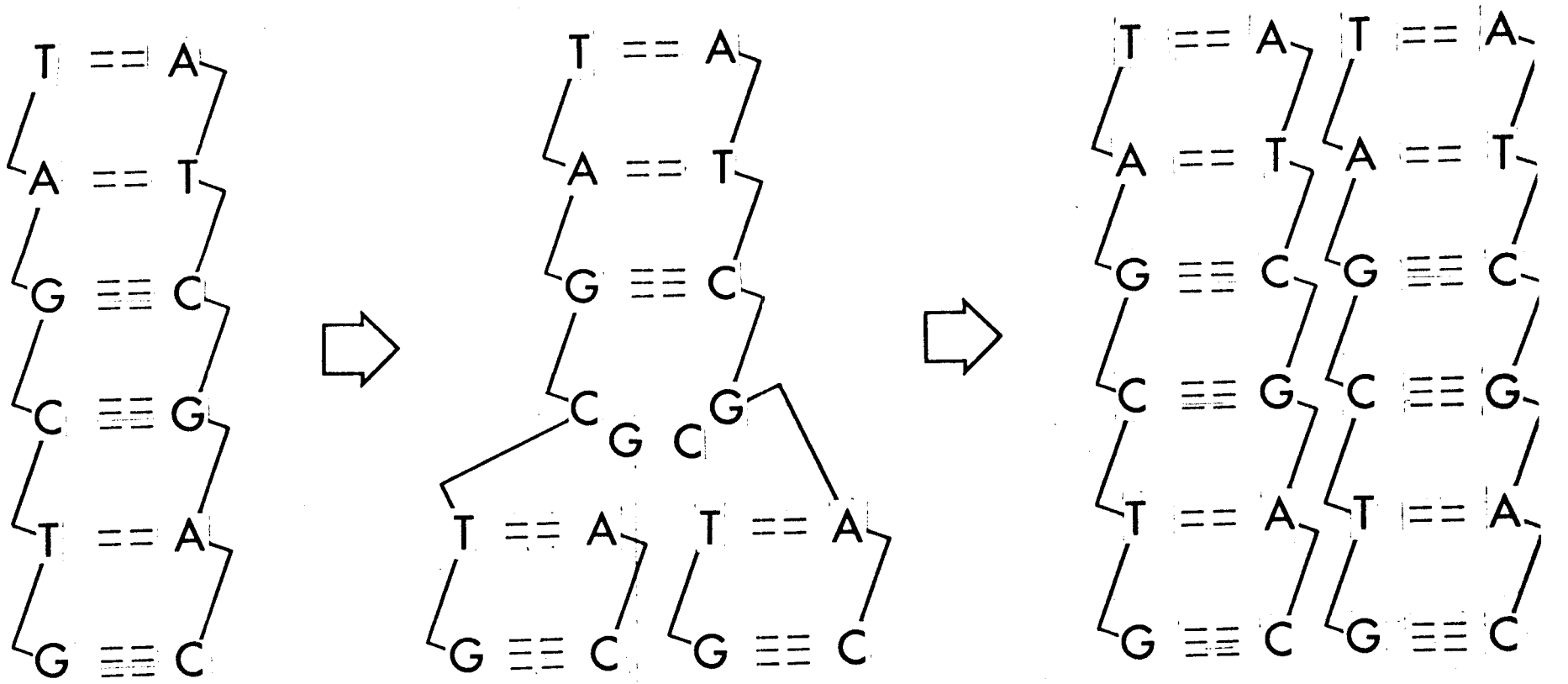
*Hb -human hemoglobin, TMV -tobacco mosaic virus protein, TS -tryptophan synthetase of E. coli, L -beta lactoglobulin, bovine, AP -alkaline phosphatase of E. coli, Cyt c -human cytochrome c.**Spontaneous changes in TMV.

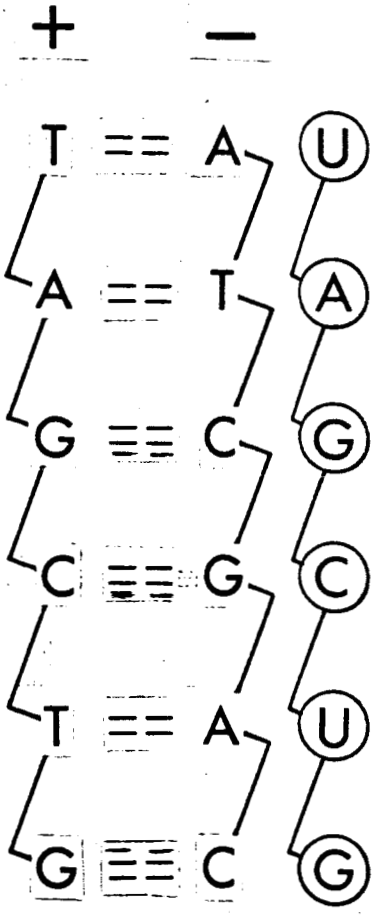
Figure 1. Illustrating the structure of part of a DNA molecule. The dotted lines indicate hydrogen bonds that hold the two chains together.

Figure 2. Method of formation of two double strands of DNA (c) from a parent double strand (a). In (b) the complementary Watson-Crick pairing that accompanies the replicating procedure is in progress.

Figure 3. Formation of RNA (...UAGCUG...) by modified complementation with one strand of DNA.


DEOXY RIBOSE
Thymine
Adenine
Phosphate
DEOXY RIBOSE
Cytosine
Guanine
Phosphate
DEOXY RIBOSE
Guanine
Cytosine
Phosphate
DEOXY RIBOSE
Adenine
Thymine


T == A
A == T
G == C
C == G
T == A
G == C
T == A
A == T
G == C
C G
G C
T == A
G == C
T == A
G == C
T == A
A == T
G == C
C == G
T == A
G == C
T == A
A == T
G == C
C == G
T == A
G == C


+
—
T == A
U
A == T
A
G ≡≡ C
G
C ≡≡ G
C
T == A
U
G ≡≡ C
G