THE GENETIC CODE

D. J. NOLTE, D.SC. (RAND), *Department of Zoology, University of the Witwatersrand, Johannesburg*

Since the elucidation of the chemical constitution of the hereditary factors or genes, it has become evident that genes can perform their functions only through the intermediation of systems of coding and decoding. These systems are composed of two kinds of nucleic acid, viz. deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). In short, chromosomal DNA directs the synthesis of messenger RNA which in turn directs the synthesis of proteins, with the assistance of adaptor molecules called transfer RNA.

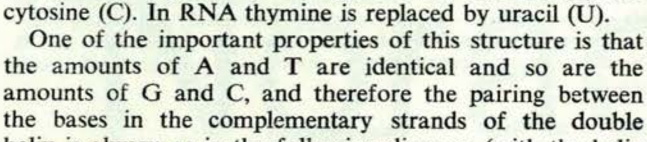
THE CODING SYSTEM

When it was discovered that hereditary traits of pneumococci could be transmitted from one strain to a genetically different strain by means of purified DNA derived from the donor,[1] it led to the establishment of the vital fact that DNA is the carrier of genetic information, and laid a cornerstone of molecular biology. This discovery provided a tremendous impetus to the study of DNA, which was known to be an integral part of the chromosomes in all cells of all organisms.

DNA

This nucleic acid occurs in chromosomes which are composed of nucleo-proteins, i.e. DNA in association with about equal amounts of basic proteins or histones. The Watson-Crick model[16] for the physical structure of the DNA molecule is based especially on X-ray diffraction experiments—a double helical chain (two strands twisted round each other like the strands of a rope) with the backbone of each of the strands composed of molecules of deoxyribose sugar linked with one another by phosphate bonds. Inside the helix a nitrogenous base is linked to each sugar molecule and the two opposite bases are cross-linked by hydrogen bonds. DNA is thus a polynucleotide; a nucleotide being composed of a base, deoxyribose and phosphate. The bases are the two purines adenine (A) and guanine (G), and the two pyrimidines thymine (T) and cytosine (C). In RNA thymine is replaced by uracil (U).

One of the important properties of this structure is that the amounts of A and T are identical and so are the amounts of G and C, and therefore the pairing between the bases in the complementary strands of the double helix is always as in the following diagram (with the helix flattened out).

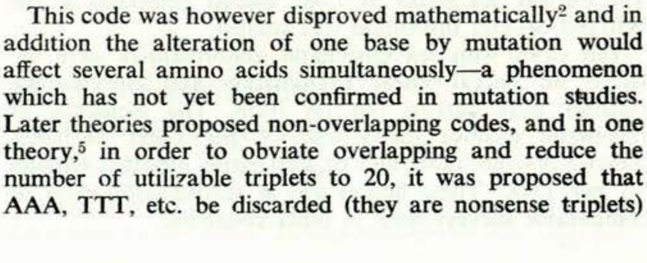


Reproduction of this molecule is probably by the separation of the two strands, each then building up a new complementary partner from metabolic substrates through the specific pairing of bases. Thus each DNA strand is a template for the production of a second strand with a complementary set of bases so that in cell division the daughter cells receive identical copies of DNA. It is generally believed that during mutation of genetic material one method of deviation is by wrong copying, i.e. a particular base is replaced by another, so that in replication this forms its own complement.

Because the amounts of DNA are constant in different cells of the same organism but vary greatly for different species, and also because the amounts A+T and G+C differ for different organisms, it appears that not only is DNA the carrier of genetic information, but that the kind of information is variable owing to variable amounts of bases. Specificity of DNA lies in the arrangement of the 4 bases, i.e. it carries coded information in the sequence of bases along the chain.

The Code

It is generally believed that the gene functions by controlling a specific unit process in the interrelated chains of biochemical reactions. The prime function of the gene thus lies in determining the specificity of some particular enzyme which will control such a process. Enzymes are polypeptides or proteins which are composed of various combinations and permutations of amino acids linked together, and of these amino acids 20 different kinds are usually found in natural polypeptides; the sequence of amino acids gives the protein its individual character. If DNA has a sequence of nucleotides or base pairs along the chain, which provides genetic information to determine the sequence of amino acids in a polypeptide chain, it is evident that the 4 bases must be fitted into a code to represent the 20 amino acids. To satisfy this condition triplet codes were suggested as forming the most probable coding system. A triplet code has three adjacent base pairs to represent a particular amino acid and such a triplet is called a codon. Now the difficulty is that for 4 bases there are $4^3=64$ different triplets to represent 20 different amino acids. To overcome this excess some very interesting codes have been proposed. One of the most popular for a while was the overlapping code,[7] as follows:



This code was however disproved mathematically[2] and in addition the alteration of one base by mutation would affect several amino acids simultaneously—a phenomenon which has not yet been confirmed in mutation studies. Later theories proposed non-overlapping codes, and in one theory,[5] in order to obviate overlapping and reduce the number of utilizable triplets to 20, it was proposed that AAA, TTT, etc. be discarded (they are nonsense triplets) and that the remainder be grouped into 20 sets of 3 each (each set being cyclic permutations of the others, e.g. of ATG, GAT, TGA, only one makes sense). One solution in this theory gives the following 20 sequences:

```
                                    A
                      A     A
        A     A                     T
A   T             G   T     T   C
        T     T                     G
                      G     G
                                    C
```

This would give a coded DNA chain with no overlapping between adjacent triplets; however, this marvellous coincidence has not been confirmed (see later). Proof that the coding ratio is 3 (or a multiple) was provided by a group of Cambridge workers,[4] who suggest a non-overlapping code of 3 bases per amino acid, with the sequence reading from a fixed starting point. With the existence of 64 possible triplets, they point out that the code is probably degenerate, i.e. in general more than one triplet codes for each amino acid.

A major breakthrough came when it was discovered that non-living systems, prepared from the colon bacillus, could synthesize artificial polypeptides from artificial polymers of nucleotides. Using as intermediary the single strand RNA with the base U instead of T, the first polymer produced was poly-uracil (a polynucleotide consisting of U only) and this synthesizes the amino acid phenylalanine into polyphenylalanine. By using a particular enzyme polynucleotide phosphorylase, two groups of workers were able to synthesize and test polymers of all possible combinations of the bases A. G, U and C and to allocate various triplets or codons to various amino acids.[10-15] For example, CCC was found to code for proline, AAA to code for lysine, AAU to code for tryptophane, etc. Up to date about 50 different triplets have been reported as codons, and since some of them code for more than one amino acid it appears that the genetic code is indeed degenerate. However, it has become evident that two or more triplets coding for the same amino acid, often have two of their three bases in common, and it has been suggested[6] that a pattern is now emerging. The triplets which specify the detailed structure of protein fall into a regular pattern: the 64 combinations of 4 nucleotides taken 3 at a time, are resolved into 32 pairs. The second member of each pair is identical with the first, except that in one position a purine is replaced by the other purine or a pyrimidine by the other pyrimidine. From this pattern predictions were made of the amino acids which will be found to correspond to the remaining 19 unidentified triplets. Examples of this pattern are:

alanine	CCG	serine	CUU	lysine	AAA
	CUG		CCU		AGA

On this interpretation there would be only one codon for each of 9 amino acids, 2 for each of 10 amino acids and 3 for serine. It may be that in different organisms different codons are used for the same amino acid.

With such a system of triplet coding some indications of the size of the gene may be obtained. If a protein is composed of a string of 200 amino acids, the directing RNA string must have at least 600 bases in sequence and the double helix of chromosomal DNA the same number of pairs of bases or nucleotides.

THE DECODING SYSTEM

The genetic information carried in the DNA is passed on to the RNA, which must use this information in directing the specificity of the polypeptide chain being constructed, i.e. the sequence of bases must determine a sequence of amino acids. In other words, RNA is the component of the system for decoding the information of the gene. RNA, as the intermediary between DNA and protein, is basically similar in structure to DNA, but shows no regularities in base ratios since it is essentially a single-strand structure; the sugar is ribose of which the molecules are linked by phosphate, while a base is linked to each sugar with uracil replacing thymine. It has been known for some time that the cytoplasm of cells contains soluble and insoluble RNA but lately it has been determined that the actual copy of the gene is found in a third, relatively unstable, form.[3]

Ribosome RNA

This r-RNA forms the surface on which protein synthesis takes place. It is found in ribosomes, which are small granules attached to the endoplasmic reticulum of the cytoplasm of the cell. Ribosomes are composed of equal amounts of protein and RNA, with the latter partly in a helical structure. The overall base composition of the r-RNA is remarkably similar in a variety of organisms, and this fraction of the RNA cannot synthesize protein unless it is given instructions about the sequence of assembly of amino acids. The latter instruction is given by the next kind of RNA, and ribosome RNA thus seems to function simply as a stable surface for the decoding of information.

Messenger RNA

This m-RNA forms the template for the decoding of the information carried by the DNA. It is a single strand of polynucleotide which is catalyzed on the DNA helix by RNA-polymerase. It is not yet known whether m-RNA is a complementary of only one of the double DNA strands, or whether these strands unwind temporarily to form a composite complementary RNA strand. Messenger RNA represents the transmitted code from one gene and comprises about 1,000 nucleotides; it appears to be relatively unstable as if it is broken up when its function is completed. It leaves the nucleus and is attached to a ribosome which forms a stabilizing surface for maintaining the m-RNA strand in the correct configuration. The sequence of its bases will specify the amino acid sequence in the protein when these amino acids are linked by peptide bonds while being unloaded from the third type of RNA.

Transfer RNA

The soluble t-RNA[8] is the active carrier of amino acids from the general metabolic pool to the site of protein synthesis. Transfer RNA consists of relatively short strands of less than 70 nucleotides, and each molecule is folded back on itself like a hairpin and then twisted into a helix. The 3 bases ACC are at one free end and G at the other of all such t-RNA molecules, but at the loop at the top of the hairpin are 3 unpaired bases which comprise the codon. If there are 32 different codons then there should be the same number of t-RNA molecules. Actually such a codon will be complementary to that on the m-RNA strand, e.g. t-RNA will carry AAA when the m-RNA codon is UUU. Each soluble transfer segment collects a specific amino acid molecule by chemical linkage at the ACC end; this linkage is attained by means of specific activating transfer enzymes so that cells should contain 20 different kinds of the latter. The amino acid is then transported to the template RNA on the ribosome and is there unloaded, as follows: The t-RNA hooks on to the correct codon of the m-RNA template by means of the complementary pairing properties of its 3 central bases so that the amino acid is now suspended at the correct point in the growing protein chain, lying adjacent to another amino acid which is suspended in a similar way; the 2 amino acids are now linked by a peptide bond. In this way the polypeptide chain is assembled while the transfer RNA molecules are detached to repeat their function.

It is not yet known where t-RNA is produced; it may be replicated on the DNA helix or may be produced by the nucleolus which is a store of RNA in the nucleus. Neither are the functions of the other bases in t-RNA known; it may be that they form a chemical pattern which is recognized by the activating enzyme so that only the correct amino acid is collected. The picture of the decoding of genetic information is then the following: A t-RNA picks up a particular amino acid and carries it to the assembly line, which consists of a template m-RNA strand stabilized on a ribosome, the complementary codons of transfer, and messenger RNA pair, and the amino acid is suspended in line with others adjacent on either side to which it is linked by peptide bonds; the t-RNA molecules are then detached.

THE OPERON

A problem which arises in visualizing the coding and decoding systems is the question of when a gene, which is a section of the DNA chain, will initiate or end its cycle of functioning. An interesting theory[9] has been advanced to elucidate this problem. The synthesis of specific proteins may be induced or suppressed under the influence of external agents. Such influences could be due to chemical operations which control the rate of transfer of information from the gene to the protein and may thus affect this process at the genic and cytoplasmic levels. The new concept is that while the structural gene forms a cytoplasmic transcript in the m-RNA strand, an adjacent operator gene is present which coordinates the activity of the structural gene (perhaps several different structural genes). This coordination is achieved through the operator combining specifically and reversibly with a substance which has a complementary structure, and this combination blocks the formation of m-RNA. Such a cluster of genes, structural and operator, is called an operon.

The specific repressor is synthesized by a regulator gene which need not be adjacent to the operon. In an inducible enzyme system, for example, the repressor may combine with certain small molecules such as the substrate used for the enzyme, so that it then has no affinity for the operator resulting in the activation of the operon, i.e. m-RNA is produced. In a repressible enzyme system the repressor is inactive and is activated only by combination with certain molecules, perhaps a metabolite of the product, so that if there is sufficient of the latter the operon will be inhibited and further production of m-RNA will stop. This whole system has been demonstrated to work in the case of the enzyme galactase in the colon bacillus.

The whole complex system of enzyme or protein synthesis is vulnerable to modification at various levels. The genic or DNA code may change owing to a mutation which could result from miscopying during its reproduction, and result in a mutant gamete at the hereditary level or a mutant cell at the somatic level; the latter could initiate the growth of a mosaic patch of cells. A pleiotropic effect may follow the mutation of an operator gene and, by affecting several structural genes of the operon, cause a syndrome of effects in an offspring, if it happened in a gamete. A phenocopy is an abnormality which resembles the effects of a mutation, but is not heritable since the specific gene has not mutated, but has only had its function modified; this may well be due to particular external factors involving the cytoplasmic functioning of the operon and its regulator gene.

REFERENCES

1. Avery, O. T., MacLeod, C. M. and McCarty, M. (1944): J. Exp. Med., **79**, 137.
2. Brenner, S. (1957): Proc. Nat. Acad. Sci. (Wash.), **43**, 687.
3. Brenner, S., Jacob, F. and Meselson, M. (1961): Nature (Lond.), **190**, 576.
4. Crick, F. H. C., Barnett, L., Brenner, S. and Watts-Tobin, R. J. (1961): *Ibid.*, **192**, 576.
5. Crick, F. H. C., Griffith, J. S. and Orgel, L. E. (1957): Proc. Nat. Acad. Sci. (Wash.), **43**, 416.
6. Eck, R. V. (1963): Science, **140**, 477.
7. Gamow, G. (1954): Nature (Lond.), **173**, 318.
8. Hoagland, M. B. (1959): Sci. Amer., **201**, 55.
9. Jacob F. and Monod, J. (1961): J. Molec. Biol., **3**, 318.
10. Jones, O. W. and Nirenberg, M. W. (1962): Proc. Nat. Acad. Sci. (Wash.), **48**, 2115.
11. Lengyel, P., Speyer, J. F. and Ochoa, S. (1961): *Ibid.*, **47**, 1936.
12. Lengyel, P., Speyer, J. F., Basilio, C. and Ochoa, S. (1962): *Ibid.*, **48**, 282.
13. Nirenberg, M. W. and Matthei, J. H. (1961): *Ibid.*, **47**, 1588.
14. Speyer, J. F., Lengyel, P., Basilio, C. and Ochoa, S. (1962): *Ibid.*, **48**, 63.
15. *Idem* (1962): *Ibid.*, **48**, 441.
16. Watson, J. D. and Crick, F. H. C. (1953): Nature (Lond.), **171**, 737 and 964.