Towards the genetic code

Francis Crick

The inherited master plans controlling every living organism are written on the genetic material in each cell. These plans are coded instructions to the cell for making proteins. The recent breaking of the code opens a new era in biology

How do living organisms function, and in particular how do they reproduce themselves? A bacterial cell, less than a thousandth of an inch in diameter, can carry out, in a controlled way, perhaps a few thousand different chemical reactions. How does the genetic message contained in the cell control this complex activity? When such a cell divides it produces two daughter cells which are very similar to itself—the genetic message passed on to each is identical. How is this precise copying process carried out?

Such questions are fundamental. The ability to feed on the environment and build complexity, and the ability to maintain and pass on these built-up complex patterns, are the essential features of life. When we can answer them—in terms of the patterns and interactions of atoms and molecules—we will understand and perhaps be able to control in some way the most basic mechanisms of life processes.

In the last few years we have come very close to finding these answers. Building on the firm foundations of modern physics, chemistry and genetics, it has proved possible to approach living things from the atomic level upwards. This new approach, now usually described by the general term 'molecular biology,' has made remarkable progress in the last decade or so. For example, we can now see in broad outline, and a good deal of detail, what the genetic instructions are made of, how they are passed on from cell to cell, and how they carry out their vital task of manufacturing proteins. Further, and perhaps more remarkable, this new approach has led to two fundamental generalisations. First, we now know that the chemistry of the basic biological processes is remarkably similar throughout the whole of Nature. For example, the genetic material of a bacterial cell is very similar to our own, and the proteins that it makes almost identical to ours in overall composition. Secondly, we now realise that, though superficially complicated, the crucial mechanisms of living things are specified at the atomic level, and in a rather simple way.

Structure of proteins

Two great families of molecules control the key functions of the living cell. They are the proteins and the nucleic acids. It is convenient to deal with the proteins first.

The main function of proteins is to act as 'enzymes'—the highly specialised catalysts which speed up the chemical reactions in the cell. Under the mild conditions of temperature and acidity within a cell most of these reactions would take place only extremely slowly if these catalysts were not present. Each catalyst acts in a highly specific way on a particular chemical reaction. Thus if the cell can produce the correct set of proteins much of the rest will follow. On a larger scale proteins make up much of the structure of our living machinery; they are the stuff of skin, hair, muscle, blood vessels, and internal organs.

Proteins are large molecules, typically containing thousands of atoms (*see Fig. 3*). But though large, their basic chemical structure is remarkably simple. A protein usually consists of a single polypeptide chain; that is, a long chain with a backbone having a regular repeating structure to which different side-groups are attached at regular intervals. The structural units of this chain—the monomers which are joined together to form the protein polymer—are amino acids of which there are 20 different kinds. A particular protein, which may be several hundred amino acid units long, has the amino acid residues arranged in a very definite sequence. Since the classic work of Frederick Sanger on the sequence of insulin, we now know in full the amino acid sequences of several proteins.

This sequence is the so-called ' first order structure ' of the protein. Once formed, this chain folds on itself to form the ' second order structure,' often a regular helix like that which can be seen in parts of the myoglobin model shown in Fig. 3. This helix also folds on itself in a precise but complicated way so that each protein has an intricate three-dimensional shape peculiar to itself and different from that of any other protein. It is this so-called ' third order structure ' which allows each protein to carry out its special job. Thus proteins, as a family, are delicate, subtle and versatile. But behind this complexity their basic chemical structure—the linear sequence of amino acid units—is rather simple, which means that they can be put together by a relatively simple process.

Structure of nucleic acids

How *is* this sequence put together? How does the cell contain the information needed to put the 20 amino acids into the correct order for each of the thousands of proteins it may make? This appears to be the main function of the genetic material. The genes, the units of genetic function, are arranged in a linear order along the chromosomes, which in higher organisms reside in the nucleus of the cell. Each particular gene probably contains the instructions to make one particular protein. This is the ' one gene—one enzyme' hypothesis.

From a great deal of evidence we now believe that genes are made from the other great family of biological molecules, the nucleic acids. There are two kinds of nucleic acid, closely related to each other, called DNA (for deoxyribonucleic acid) and RNA (ribonucleic acid). The genetic material is usually DNA, though for some small viruses, such as polio, it is RNA. Most of the RNA in cells has other, though related, functions.

DNA is also a polymer, and a very

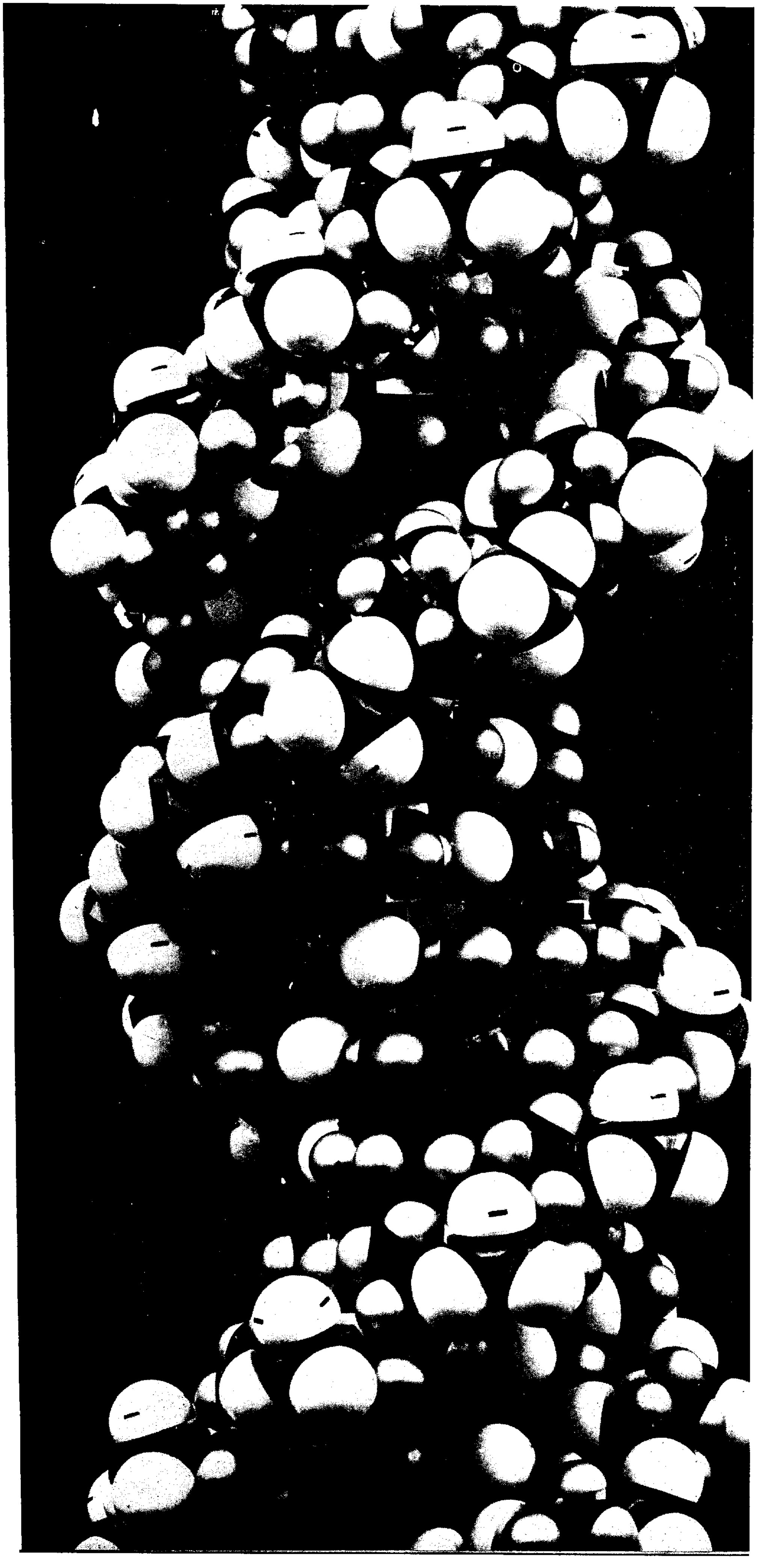


Fig 1 This model of a length of DNA—the genetic material—shows double helical structure proposed by Crick and Watson. (Courtesy of M. H. Wilkins)

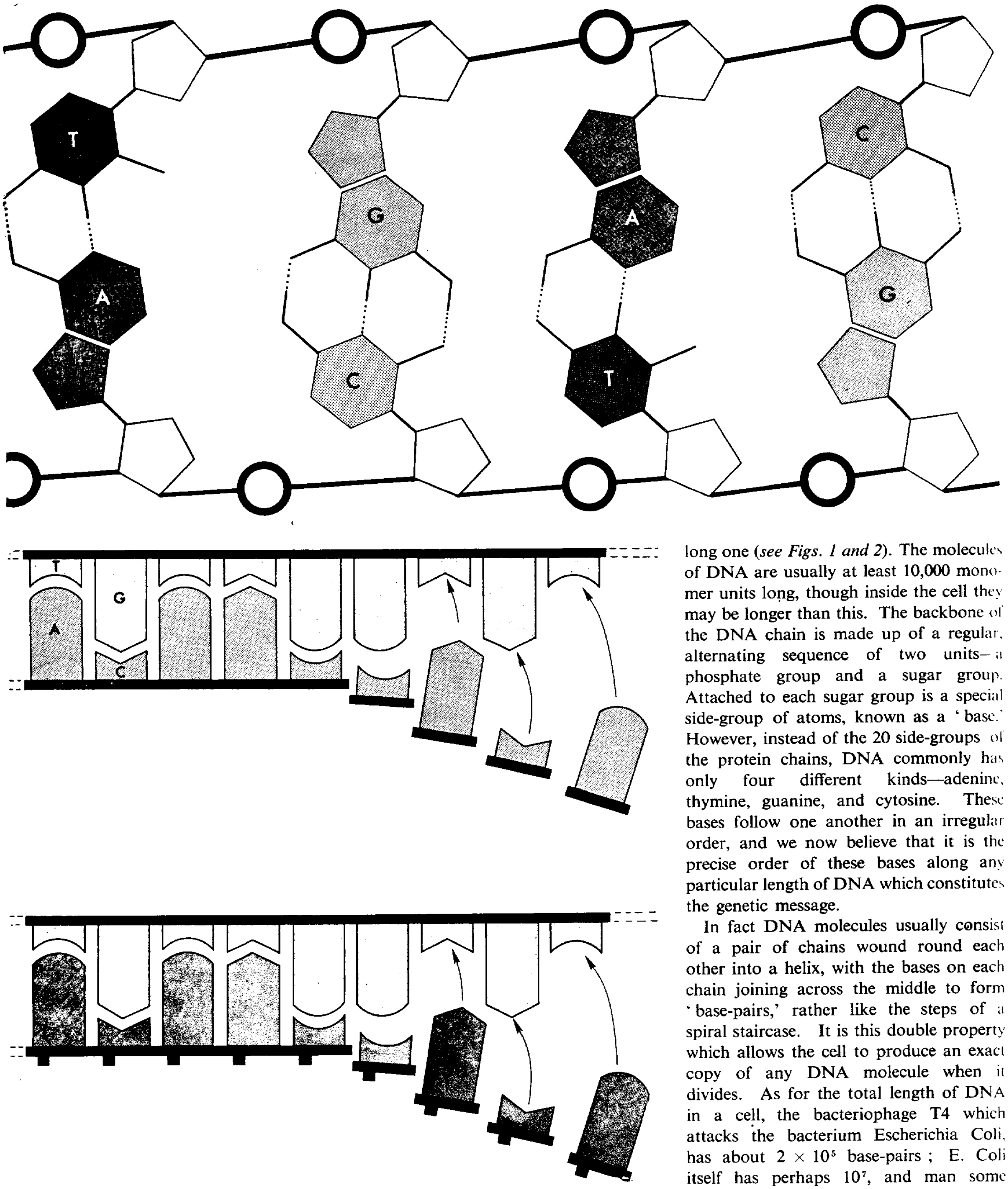


Fig 2 Top drawing shows basic structure of DNA—a double backbone of sugar and phosphate groups joined by base-pairs A-T and G-C (adenine, thymine; guanine, cytosine). Lower drawings show (top) half DNA molecule acting as template for forming other half, and (bottom) as template for forming related RNA

long one (*see Figs. 1 and 2*). The molecules of DNA are usually at least 10,000 monomer units long, though inside the cell they may be longer than this. The backbone of the DNA chain is made up of a regular, alternating sequence of two units—a phosphate group and a sugar group. Attached to each sugar group is a special side-group of atoms, known as a 'base.' However, instead of the 20 side-groups of the protein chains, DNA commonly has only four different kinds—adenine, thymine, guanine, and cytosine. These bases follow one another in an irregular order, and we now believe that it is the precise order of these bases along any particular length of DNA which constitutes the genetic message.

In fact DNA molecules usually consist of a pair of chains wound round each other into a helix, with the bases on each chain joining across the middle to form 'base-pairs,' rather like the steps of a spiral staircase. It is this double property which allows the cell to produce an exact copy of any DNA molecule when it divides. As for the total length of DNA in a cell, the bacteriophage T4 which attacks the bacterium Escherichia Coli, has about 2 × 10^5 base-pairs ; E. Coli itself has perhaps 10^7, and man some small multiple of 10^9 base-pairs in each cell—enough for over a million genes if each gene were a few thousand base-pairs long. When uncoiled, the DNA from all the cells in a human body would reach across the solar system.

Thus on this picture each gene is in fact a part of a DNA molecule—perhaps 1,000 base-pairs long—and the precise order of the four kinds of base-pairs along this length determines in some way the precise order of the 20 amino acids of the protein controlled by this gene. How this is done has come to be known as the 'coding problem.'

The coding problem

The first detailed ideas about gene-protein coding were put forward by Gamow, the astrophysicist, in 1954. Since then there has been a series of attempts, all unsuccessful, to solve the problem from a theoretical angle. In fact a year ago it looked as if the problem had got bogged down. But recently there have been dramatic developments, and it now seems likely that the code will be largely solved within a short time.

The basic difficulty in solving the code has been that while, in favourable circumstances, one can determine the amino acid sequence of a protein, it is not yet technically possible to find the base sequence of a particular piece of DNA, so the problem has to be attacked by indirect methods.

The first question we can ask is how many bases are needed to determine one amino acid. If only a pair of bases were used we should have only 4 × 4 = 16 combinations, whereas there are at least 20 kinds of amino acids in proteins. Thus the minimum number of bases needed is 3. There are 4 × 4 × 4 = 64 possible triplets, and it is not obvious how they should be allocated to the 20 amino acids. For example, each amino acid might have just one triplet and the other 44 triplets might be 'nonsense'—that is, have some other function. Alternatively the code might be 'degenerate'—that is, several triplets might stand for each amino acid. In either case one might expect one or more triplets to stand for a 'space,' or even that there would be separate triplets for 'begin chain' and 'end chain.'

The code suggested by Gamow was an 'over-lapping' code. This is illustrated in Fig 4 which shows an over-lapping triplet code. As can be seen, bases 1, 2 and 3 code the first amino acid, bases 2, 3 and 4 the second, and so on. It is easy to see that with such a code some sequences of

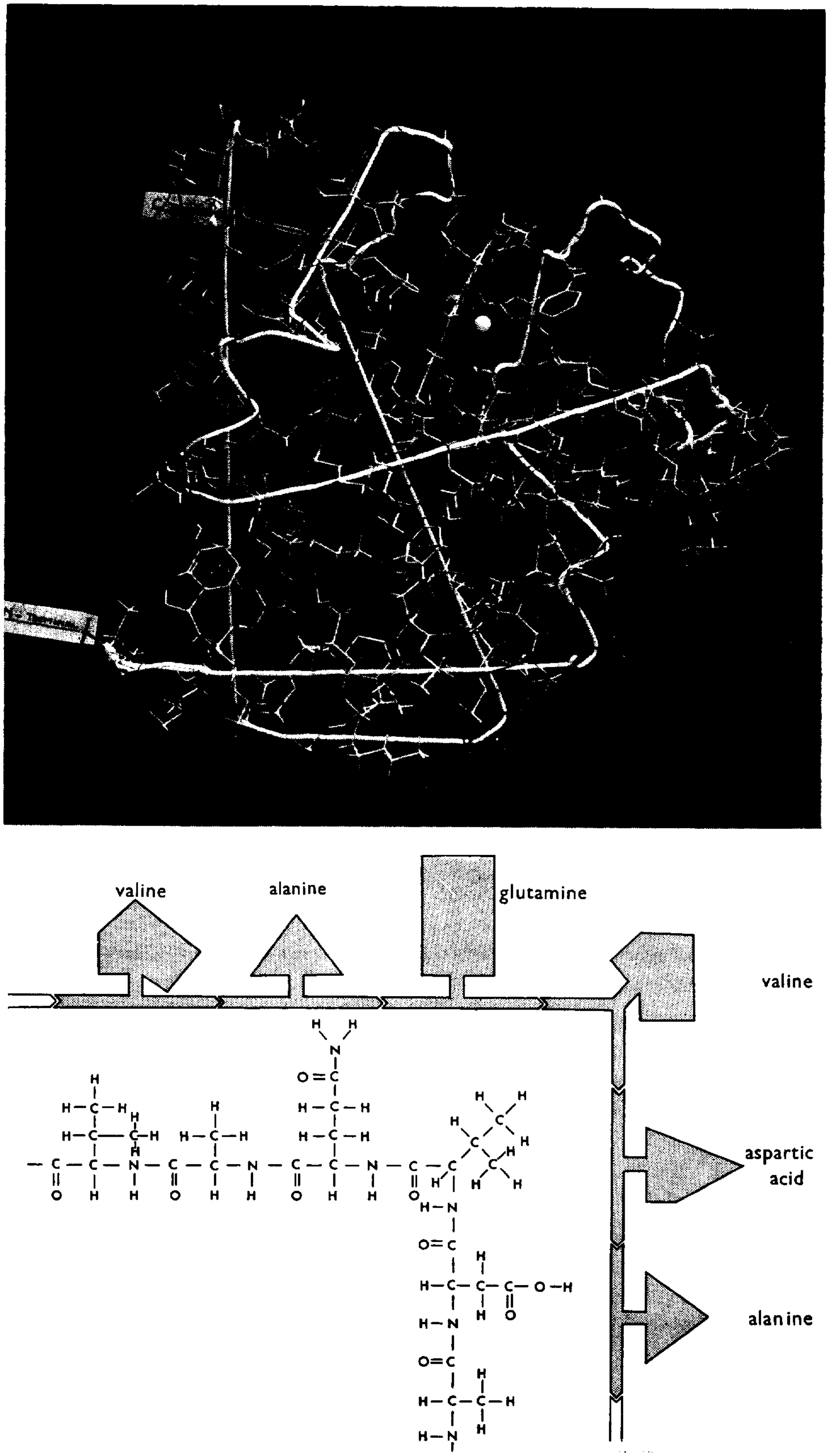


Fig 3 Top picture shows three dimensional structure of the protein myoglobin; cord shows main directions of folded amino acid chain. (Courtesy of J. Kendrew.) Diagram below shows typical short length of amino acid sequence forming all proteins; joined by peptide links, 20 different amino acids may occur in chain

amino acids cannot be coded, so that these should be restrictions on the sequence actually observed. It was soon shown that the actual family of codes proposed by Gamow was incorrect, since they could not code known sequences. Later, Sidney Brenner, by an ingenious argument, showed that if the code were 'universal' (the same in all organisms) so that all the experimental data could be lumped together, then *no* simple overlapping triplet code was possible.

Recently more direct evidence has confirmed this. In an overlapping code a change of one base will, in general, alter *three* adjacent amino acids. Such changes or 'mutations' may have occurred 'spontaneously,' or they may be produced by chemical means. One way of artificially changing the sequence is to treat the genetic material with nitrous acid, which deaminates the bases. The base cytosine, for example, is changed by this method into the base uracil.

This method has been used on the plant virus Tobacco Mosaic Virus (TMV), whose genetic material is not DNA, but the related RNA. The rest of the virus is built up from a rather simple protein consisting of a single polypeptide chain of 158 amino acids, whose precise amino acid sequence has been worked out by two groups of workers, at Berkeley and at Tubingen. After the virus (or its RNA) has been treated with nitrous acid it is used to infect the plant. The modified virus multiplies inside the plant cells so that a large amount of new virus is produced. The protein of the virus is then examined to see how its amino acid sequence has been altered. Most of this work has been done by Dr H. G. Wittman of Tubingen, who found that in no case were two or three adjacent amino acids changed. The typical alteration was to a single amino acid. It thus seems virtually certain that the code is not of the overlapping type.

If the code is not overlapping a new problem arises. How does one tell where one triplet starts and the next begins? For example, if the sequence in the middle of a gene is, say,

. . . . CATCATCAT . . . .

(where C stands for cytosine, A for ademine and T for thymine) is this to be read as

. . . . CAT CAT CAT . . . .

or

. . . . C ATC ATC AT . . . ?

Various ingenious solutions to this problem have been suggested, but we now believe that none of these is correct. It seems likely that the message is read by starting from a fixed starting point and going along three at a time from there.

Experimental evidence

The evidence for this comes from genetic work carried out at Cambridge by my colleagues, Mrs Leslie Barnett, Dr Sydney Brenner, Dr Richard Watts-Tobin, and myself. The system we used was a particular gene, the B cistron of the r_{II} gene, of the bacteriophage T4, which attacks the bacterium Escherichia Coli. This is the gene so brilliantly explored by Dr Seymour Benzer. The choice of this gene was dictated by the fact that one can study rather rare events and that the experiments can be done very quickly. One can find rare events because it is possible to handle very large numbers of individuals, and because special techniques allow one to pick out the virus one wants from among a large number of them one doesn't want.

The basic operation in our experiments was the genetic 'cross.' Two different variations of the T4 phage are allowed to infect one bacterial cell at the same time (*see Fig. 5a*). After 25 minutes the cell bursts open and about a 100 new viruses emerge. Some of these will be like the first parent, and some like the second parent, but in addition there will be some having mixed properties, that is with some characteristics of the first parent and some of the second.

Consider, for example, a gene with a defect in it at the point A (*see Fig. 5b*), and another virus with a defect at point B in the same gene. When these are crossed together a few of the progeny will have both defects, at A and at B, and a few will have no defect.

It turns out that the nearer A or B are together on the gene, the more rarely will the 'recombination' take place. Moreover, if the mistakes A and B are at the 'same' place, a good copy will *never* be produced. It is by this basic procedure that Benzer was able to map the r_{II} gene and show that it had many different sites arranged in a linear array.

Our genetic studies were made with a mutant—an altered variation of the gene—produced by the action of proflavin. There is indirect evidence that proflavin makes mutants by adding or deleting a base, rather than by changing one. When

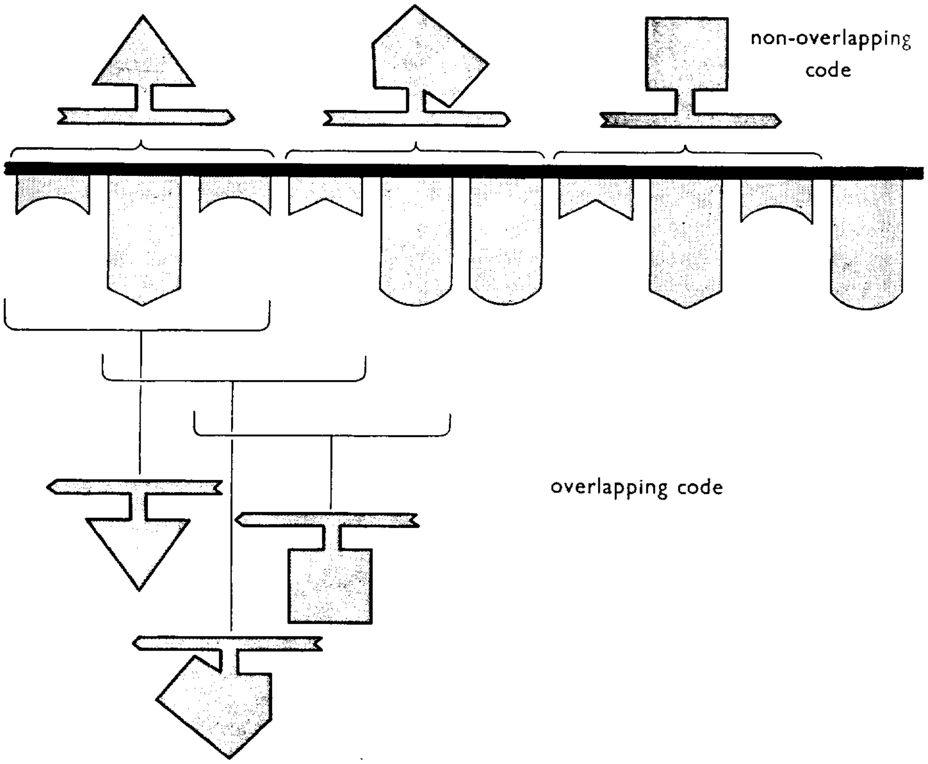


Fig 4 Triplets of DNA bases form specific code for amino acids. Early ideas favoured overlapping code (code is now known to be non-overlapping)

we grew a stock of our altered virus, occasionally a virus appeared in which this gene was functioning again. We could pick up this rare event (perhaps one virus per million) because of the powerful selective techniques we could use to look for it.

We then found, using the technique of genetic mapping, that this second alteration was not usually a correction at the original site, but was due to a further change at a nearby site. We found that either of these alterations, *by themselves*, could remove the function of the gene, but that when both alterations were together in the same gene the function was restored, though not completely. Thus, using Fig. 5b for explanation, defect A alone, or defect B alone were non-functional; but with both defects, A and B, the gene worked fairly well.

This suggested the following explanation (*see Fig. 6*). If the genetic message is read in groups of three from one end, then the addition of a base near the beginning of the message will alter *all* the following triplets. This explains why such a gene has no function. However, let us suppose that the *second* defect is due to the *removal* of a base. Then although the few triplets between the two alterations will be changed,

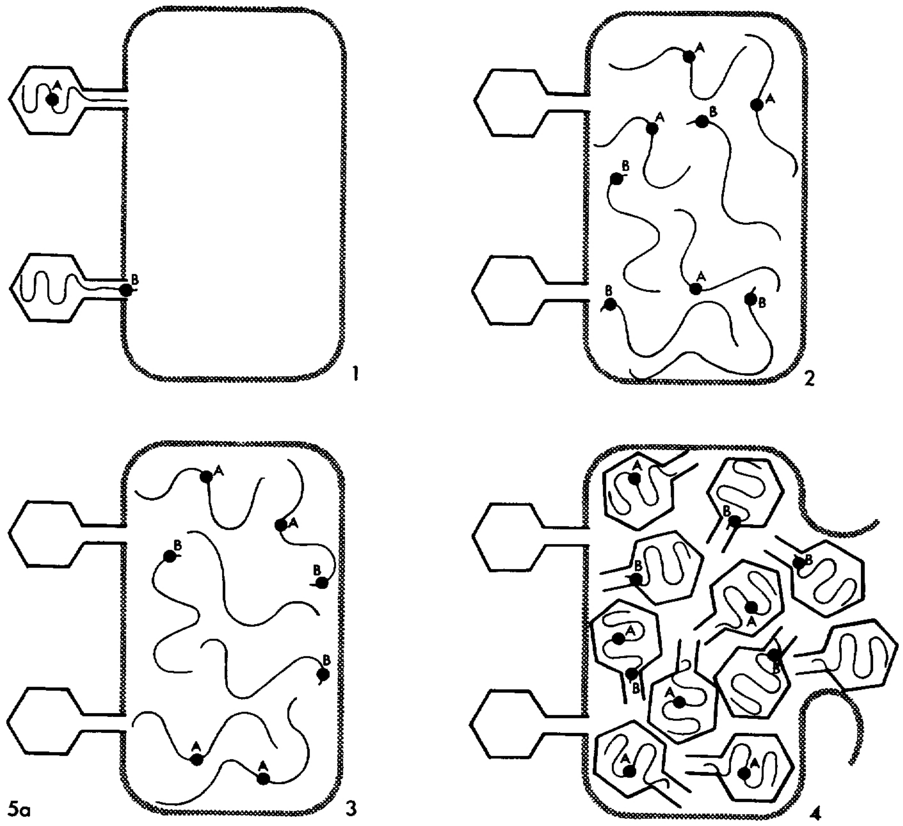
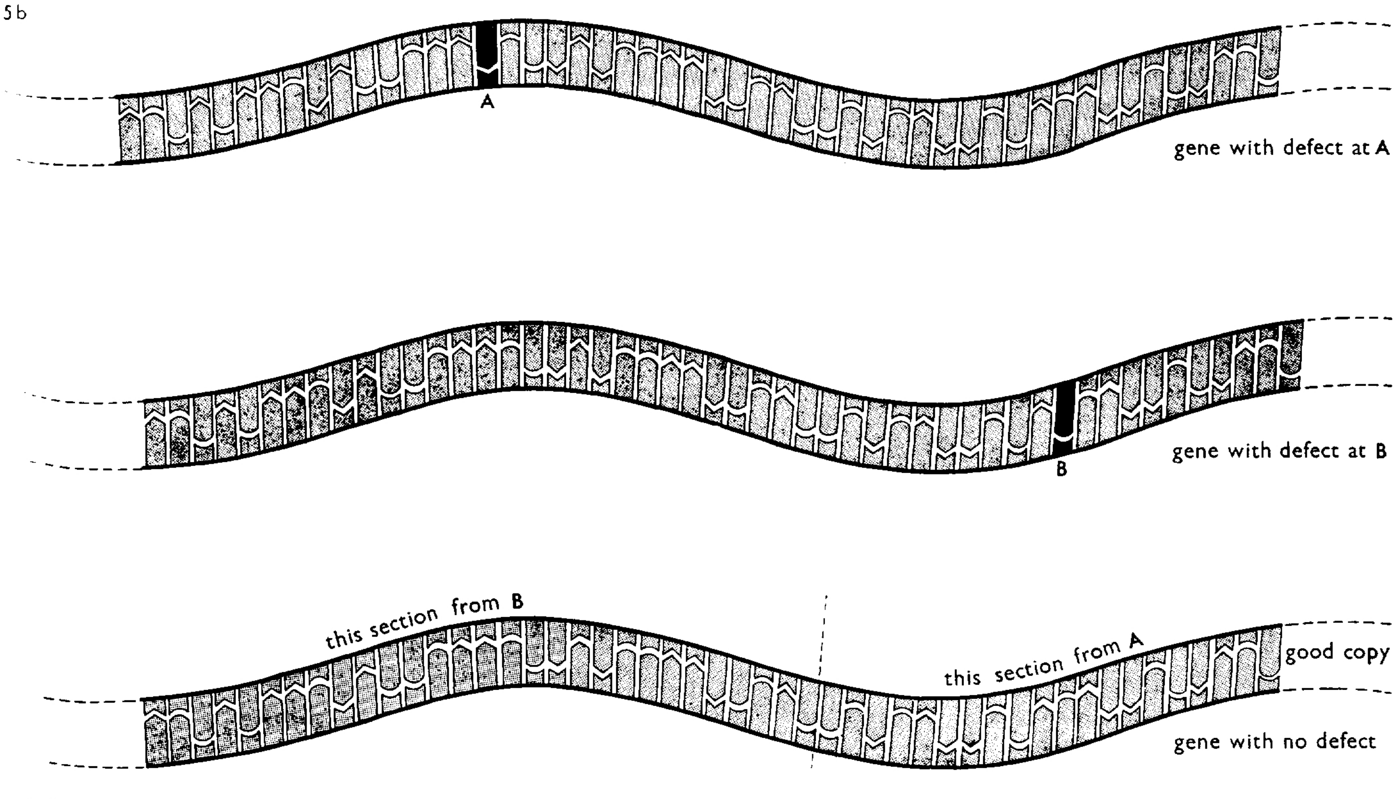


Fig 5 Code-solving technique involved infection of bacteria by viruses with defects (A, B in 5a) on same gene. 5b shows part of genes enlarged—see text



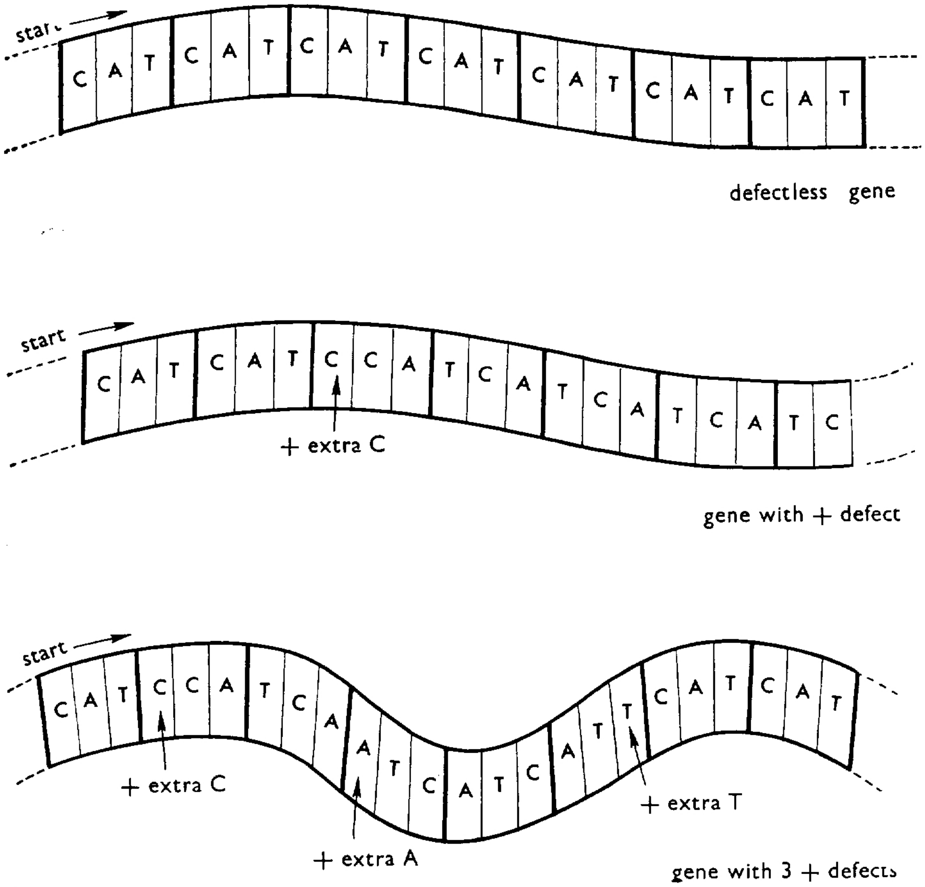

start
C A T C A T C A T C A T C A T C A T C A T
defectless gene
start
C A T C A T C C A T C A T C A T C A T C
+ extra C
gene with + defect
start
C A T C C A T C A A T C A T C A T T C A T C A T
+ extra C
+ extra A
+ extra T
gene with 3 + defects

Fig 6 Top section of DNA has no defects; triplet code (simplified here) is read from start point at left. In centre section, added base disrupts code to right. With three added bases (bottom) code eventually becomes correct

the rest of the message will be restored to its original meaning. This explains why the gene works, and also why it is not exactly the same as the original version.

In all, by these techniques, we were able to produce about 80 independent mutants within this region of the gene. We could allocate them to two classes, which we called + and −, by seeing which pairs had a workable gene. One can think of the + class as having an extra base, and the − class as having one too few. By genetic methods we made combinations of the type (+ with +) or (− with −), and, as expected, these never had any function.

So far, so good. But what if, by genetic methods, we put *three* known alterations into one gene? We made the remarkable discovery that such a triple mutant was functional.

To see why this is surprising—surprising, that is, unless one had the idea beforehand—consider this in more detail. Let us call the defects X, Y and Z. We know these are all +, say, because with another defect, say Q, which we have called −, they give a working gene. That is (X + Q), (Y + Q) and (Z + Q) all have some function. However, neither X nor Y nor Z alone has any function, nor does (X + Y) or (Y + Z) or (Z + X). But when we put together (X + Y + Z) the gene, containing all three defects, works fairly well. Looking at Fig. 6 one can see that this fits in very well with our theory. Although the region between the added bases is altered, the rest of the gene is now restored to its original meaning.

This result shows clearly that the ‘coding ratio’—the number of bases which stands for one amino acid—is either 3 or a multiple of 3. The reason that it might be 6, for example, is that it is just possible that our original mutant had *two* bases changed. If this were so all the subsequent changes would have involved an even number of bases or we should not have picked them up. However, we have subsidiary evidence, reported in our recent paper to *Nature* (December 30th, 1961) to suggest that the correct number is in fact 3, though 6 or 9 cannot be completely ruled out.

Our results also suggest that the code is ‘degenerate’—that there are not just 20 triplets which stand for amino acids and 44 which do not. If this were so we should not get combinations of the type (+ with −) to work when separated by the rather large distance we actually observe. However, this is less certain than the rest of our conclusions.

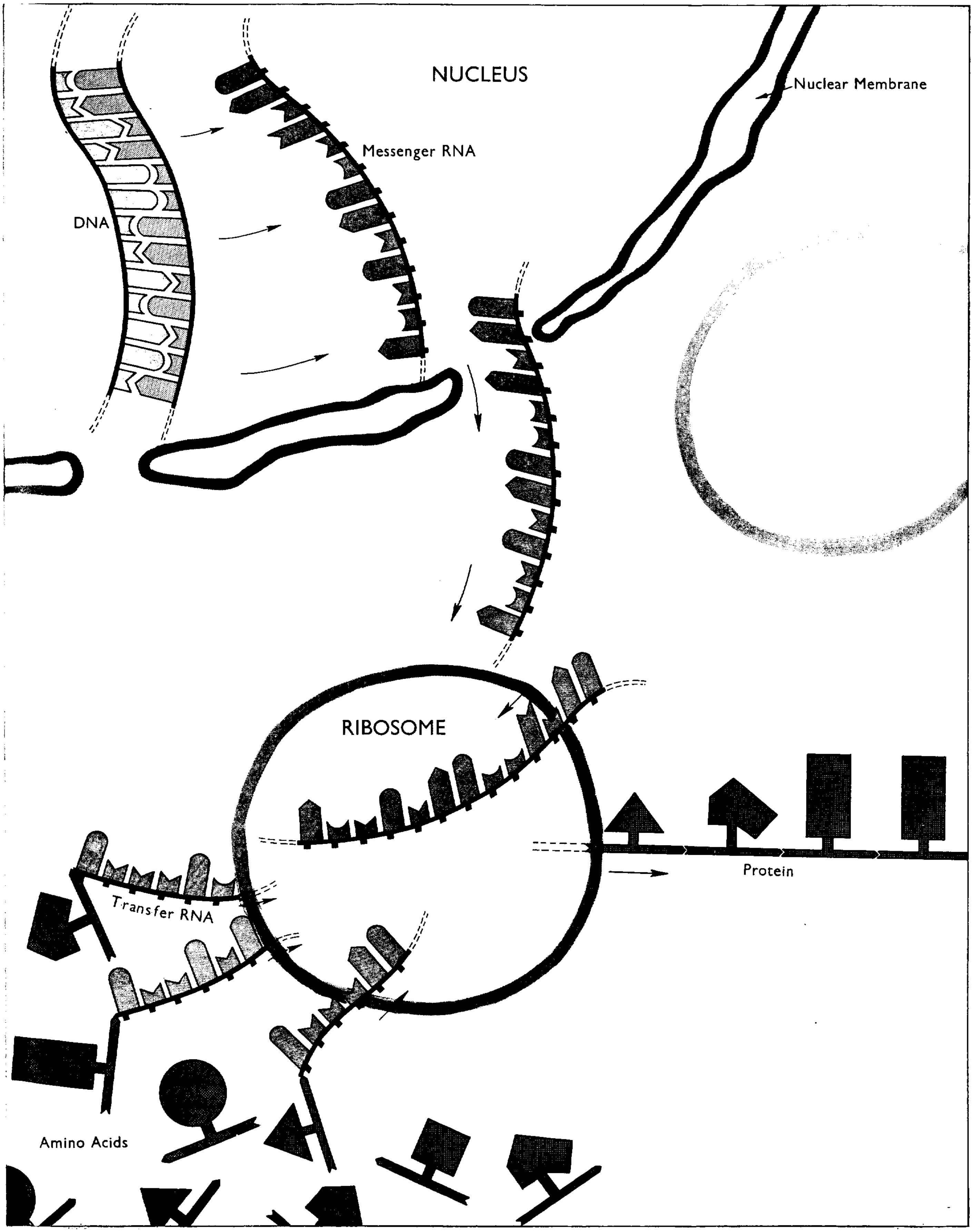
Although our genetic results show the general nature of the genetic code it would be impossible, or at least very difficult, to obtain the *details* of the code by genetic methods alone. The real breakthrough in the coding problem is due to two scientists in America who have hit on a way of doing this by biochemical methods. To understand their approach we shall have to consider the actual process of protein synthesis in more detail.

It is believed that most protein synthesis in the cell takes place not in the nucleus of the cell, where the genetic material is stored, but in the cytoplasm outside the nucleus. The actual sites are believed to be small, almost spherical particles, called ribosomes, which are roughly half RNA and half protein (*see cover picture*).

How is the genetic message transmitted to the ribosomes? At one time we used to think that the main RNA of the ribosomes contained the messenger, but recent work has suggested that there is a special RNA, now called ‘*messenger RNA*’ which is synthesised in the nucleus, as a single-stranded copy of the DNA, and which goes into the ribosomes where it acts as the actual ‘template’ for protein synthesis (*see Fig. 7*). This RNA is believed to turn over rather rapidly, at least in bacteria. This is a version of the classic slogan “DNA makes RNA, and RNA makes protein.”

How does the messenger RNA get the amino acids into the correct order? The amino acids cannot easily, by themselves, ‘recognise’ the correct base sequence of the messenger RNA. We believe this is done by each amino acid being provided with an ‘adaptor.’ A special activating enzyme, which *can* recognise the amino acid

Fig 7 Described in text, protein synthesis in cell involves DNA, messenger and transfer RNA (plus energy groups). Transfer RNA is about 80 bases long; messenger RNA 500 or more


NUCLEUS
Nuclear Membrane
Messenger RNA
DNA
RIBOSOME
Protein
Transfer RNA
Amino Acids

because of the subtle structure of the enzyme surface, is used to join the amino acid to a special adaptor, which is another species of RNA, known as *transfer RNA*. The amino acid is joined to the end of its special transfer RNA, which, we believe, then goes to the ribosomes, recognises the proper triplet of bases on the messenger RNA (by forming base-pairs between its bases and the bases of the messenger) and thus gets the amino acid into the right place. There is a special transfer RNA and a special activating enzyme for each of the 20 amino acids, and the biochemistry of this part of the process is fairly well understood, due mainly to the work of American biochemists. The energy to make the chemical link is provided by splitting ATP, the small molecule which provides the energy for many of the reactions of the cell.

It is possible to extract ribosomes from broken cells and by combining them with a 'soup'—consisting of transfer RNA and numerous enzymes, to which ATP, GTP, and amino acids are added—produce limited protein synthesis in a cell-free system.

Solving the code 'letters'

Recently two scientists in America, Nirenberg and Matthaei, were trying to stimulate protein synthesis by adding virus RNA to it. They decided to try adding a 'synthetic' RNA, which they had synthesised by a simple enzyme system, and which contained not all four bases, but just one of them, uracil, repeated many times (uracil occurs in RNA instead of the related thymine in DNA). This material, known as poly U, when added to the cell-free system stimulated the production of polypeptide chains consisting only of the amino acid phenylalanine. Thus the code for phenylalanine is probably the triplet of bases UUU.

This result was reported at the International Biochemical Congress at Moscow last August, and it was immediately apparent that there was a very good chance that the code could be solved by further work along these lines. Although we cannot yet produce a long RNA molecule with a defined base sequence, we can obtain very short molecules, having up to three or four bases with defined sequences and these short molecules can be used to start the enzymatic synthesis of long chains. In addition we can produce polymers of known composition but of random sequence. For example, it is possible to obtain poly UC, a random co-polymer of uracil and cytosine. If this is added to the cell-free system it has been found in several laboratories that the polypeptide chains produced contain only four of the twenty amino acids, namely phenylalanine, serine, leucine and proline. Such work is now in full flood, with papers slipping round referees at a great rate. It will take a little time to find the snags in the system, and put it all on a water-tight basis, but the preliminary results—for example, those of Lengyel, Speyer and Ochoa—show without doubt that the artificial RNA's are fairly specific in the effects they produce. There is, moreover, the beginning of an encouraging agreement between this work and Wittmann's work on the nitrous acid mutants of TMV.

However, even if this approach is completely successful, so that we can say in detail which triplets correspond to which amino acid, two other aspects of the problem remain. It has still not been shown that there is a simple linear relationship between the gene and the amino acid sequence of the protein it produces. This could be done by combining fine-scale genetic mapping of a gene with studies on the alterations of the amino acid sequence of the protein produced by that gene. Unfortunately, one cannot do genetics with TMV (where we have the protein) nor do we have the proteins for the r_{II} genes of T4 (where the genetic mapping is easy). However, several other systems, both in phage and bacteria, are being developed. It would be surprising if the gene and its protein were not colinear and there is now a reasonable hope that it may be proved, for one or two cases, within the next year or so.

The answer to the other problem is less certain. Is the code universal? We know that the same set of 20 amino acids is used throughout Nature, from viruses to man, but are they always coded by the same triplets? There are several experiments which suggest that this is true, or at least largely true. However, by using cell-free systems produced from different organisms we could quickly find the answers once the synthetic RNA's are available.

With luck, then, the genetic code will be solved. What consequences will this have? The most immediate result is likely to be the confirmation of all our *general* ideas about these basic mechanisms. In scientific work it is the accumulation of *detailed* evidence of several different types, which ultimately produces belief in the correctness of an idea, however pretty. During this period we may also hope to discover, using the same simple synthetic messenger RNA, the various biochemical steps in the process, and here we may still be in for a few surprises.

A new era

The general feeling, however, is that we are coming to the end of an era in molecular biology. If the DNA structure was the end of the beginning, the discovery of Nirenberg and Matthaei is the beginning of the end. From now on we shall have to study the more intricate parts of cell biology. The most immediate problem is that of gene control. What decides whether a gene shall act or not? We know something about this in bacteria from genetic studies but practically nothing of the biochemistry of the process. Eventually this will lead us on from control mechanisms in single cells to the interactions between groups of cells, and eventually to the whole process of embriological development. On the way we shall have to learn a lot more about cell membranes, both how the active transport of molecules takes place through them, and how the surfaces of different cells interact.

The genetic code may perhaps give us some hints about that speculative and difficult problem, the Origin of Life. The great biochemical uniformity of living things certainly suggests that they had a common origin, and it is not impossible that certain features of the present system actually contain, frozen into them, the early history of its development. For example, the original DNA may only have had two bases, and coded for fewer amino acids. It may be possible to spot this from the structure of the present code, once we know the details of it.

Finally there is the possibility of synthesising genes from organic chemicals, or what is more likely, modifying genes in a controlled way, so that we can turn a non-functional gene into a functional one. At the present time the technical difficulties look overwhelming, but with a detailed knowledge of the problem involved we can at least begin to ponder how we might go about it. It will certainly not be easy.