Francis Crick

Nobel Lecture, December 11, 1962

On the Genetic Code

Part of the work covered by the Nobel citation, that on the structure and replication of DNA, has been described by Wilkins in his Nobel Lecture this year. The ideas put forward by Watson and myself on the replication of DNA have also been mentioned by Kornberg in his Nobel Lecture in 1959, covering his brilliant researches on the enzymatic synthesis of DNA in the test tube. I shall discuss here the present state of a related problem in information transfer in living material - that of the genetic code - which has long interested me, and on which my colleagues and I, among many others, have recently been doing some experimental work.

It now seems certain that the amino acid sequence of any protein is determined by the sequence of bases in some region of a particular nucleic acid molecule. Twenty different kinds of amino acid are commonly found in protein, and four main kinds of base occur in nucleic acid. The genetic code describes the way in which a sequence of twenty or more things is determined by a sequence of four things of a different type.

It is hardly necessary to stress the biological importance of the problem. It seems likely that most if not all the genetic information in any organism is carried by nucleic acid - usually by DNA, although certain small viruses use RNA as their genetic material. It is probable that much of this information is used to determine the amino acid sequence of the proteins of that organism. (Whether the genetic information has any other major function we do not yet know.) This idea is expressed by the classic slogan of Beadle: "one gene - one enzyme", or in the more sophisticated but cumbersome terminology of today: "one cistron - one polypeptide chain".

It is one of the more striking generalizations of biochemistry - which surprisingly is hardly ever mentioned in the biochemical textbooks - that the twenty amino acids and the four bases, are, with minor reservations, the same throughout Nature. As far as I am aware the presently accepted set of twenty amino acids was first drawn up by Watson and myself in the summer of 1953 in response to a letter of Gamow's.

In this lecture I shall not deal with the intimate technical details of the problem, if only for the reason that I have recently written such a review[1] which will appear shortly. Nor shall I deal with the biochemical details of messenger RNA and protein synthesis, as Watson has already spoken about these. Rather I shall ask certain general questions about the genetic code and ask how far we can now answer them.

Let us assume that the genetic code is a simple one and ask how many bases code for one amino acid? This can hardly be done by a pair of bases, as from four different things we can only form 4 x 4 = 16 different pairs, whereas we need at least twenty and probably one or two more to act as spaces or for other purposes. However, triplets of bases would give us 64 possibilities. It is convenient to have a word for a set of bases which codes one amino acid and I shall use the word "codon" for this.

This brings us to our first question. Do codons overlap? In other words, as we read along the genetic message do we find a base which is a member of two or more codons? It now seems fairly certain that codons do not overlap. If they did, the change of a single base, due to mutation, should alter two or more (adjacent) amino acids, whereas the typical change is to a single amino acid, both in the case of the "spontaneous" mutations, such as occur in the abnormal human haemoglobin or in chemically induced mutations, such as those produced by the action of nitrous acid and other chemicals on tobacco mosaic virus[2]. In all probability, therefore, codons do not overlap.

This leads us to the next problem. How is the base sequence, divided into codons? There is nothing in the backbone of the nucleic acid, which is perfectly regular, to show us how to group the bases into codons. If, for example, all the codons are triplets, then in addition to the correct reading of the message, there are two incorrect readings which we shall obtain if we do not start the grouping into sets of three at the right place. My colleagues and I[3] have recently obtained experimental evidence that each section of the genetic message is indeed read from a fixed point, probably from one end. This fits in very well with the experimental evidence, most clearly shown in the work of Dintzis[4] that

the amino acids are assembled into the polypeptide chain in a linear order, starting at the amino end of the chain.

This leads us to the next general question: the size of the codon. How many bases are there in any one codon? The same experiments to which I have just referred[3] strongly suggest that all (or almost all) codons consist of a triplet of bases, though a small multiple of three, such as six or nine, is not completely ruled out by our data. We were led to this conclusion by the study of mutations in the A and B cistrons of the r_{II} locus of bacteriophage T4. These mutations are believed to be due to the addition or subtraction of one or more bases from the genetic message. They are typically produced by acridines, and cannot be reversed by mutagens which merely change one base into another. Moreover these mutations almost always render the gene completely inactive, rather than partly so.

By testing such mutants in pairs we can assign them all without exception to one of two classes which we call + and –. For simplicity one can think of the + class as having one extra base at some point or other in the genetic message and the – class as having one too few. The crucial experiment is to put together, by genetic recombination, three mutants of the same type into one gene. That is, either (+ with + with +) or (– with – with –). Whereas a single + or a pair of them (+ with +) makes the gene completely inactive, a set of three, suitably chosen, has some activity. Detailed examination of these results show that they are exactly what we should expect if the message were read in triplets starting from one end.

We are sometimes asked what the result would be if we put four +'s in one gene. To answer this my colleagues have recently put together not merely four but six +'s. Such a combination is active as expected on our theory, although sets of four or five of them are not. We have also gone a long way to explaining the production of "minutes" as they are called. That is, combinations in which the gene is working at very low efficiency. Our detailed results fit the hypothesis that in some cases when the mechanism comes to a triplet which does not stand for an amino acid (called a "non sense" triplet) it very occasionally makes a slip and reads, say, only two bases instead of the usual three. These results also enable us to tie down the direction of reading of the genetic message, which in this case is from left to right, as the r_{II} region is conventionally drawn. We plan to write up a detailed technical account of all this work shortly. A final proof of our ideas can only be obtained by detailed studies on the alterations produced in the amino acid sequence of a protein by mutations of the type discussed here.

One further conclusion of a general nature is suggested by our results. It would appear that the number of nonsense triplets is rather low, since we only occasionally come across them. However this conclusion is less secure than our other deductions about the general nature of the genetic code.

It has not yet been shown directly that the genetic message is co-linear with its product. That is, that one end of the gene codes for the amino end of the polypeptide chain and the other for the carboxyl end, and that as one proceeds along the gene one comes in turn to the codons in between in the linear order in which the amino acids are found in the polypeptide chain. This seems highly likely, especially as it has been shown that in several systems mutations affecting the same amino acid are extremely near together on the genetic map. The experimental proof of the co-linearity of a gene and the polypeptide chain it produces may be confidently expected within the next year or so.

There is one further general question about the genetic code which we can ask at this point. Is the code universal, that is, the same in all organisms? Preliminary evidence suggests that it may well be. For example something very like rabbit haemoglobin can be synthesized using a cell-free system, part of which comes from rabbit reticulocytes and part from *Escherichia coli*[5]. This would not be very probable if the code were very different in these two organisms. However as we shall see it is now possible to test the universality of the code by more direct experiments.

In a cell in which DNA is the genetic material it is not believed that DNA itself controls protein synthesis directly. As Watson has described, it is believed that the base sequence of the DNA - probably of only one of its chains - is copied onto RNA, and that this special RNA then acts as the genetic messenger and directs the actual process of joining up the amino acids into polypeptide chains. The breakthrough in the coding problem has come from the discovery, made by Nirenberg and Matthaei[6], that one can use synthetic RNA for this purpose. In particular they found that polyuridylic acid - an RNA in which every base is uracil - will promote the synthesis of polyphenylalanine when added to a cell-free system which was already known to synthesize polypeptide chains. Thus one codon for phenylalanine appears to be the sequence UUU (where U

stands for uracil: in the same way we shall use A, G, and C for adenine, guanine, and cytosine respectively). This discovery has opened the way to a rapid although somewhat confused attack on the genetic code.

It would not be appropriate to review this work in detail here. I have discussed critically the earlier work in the review mentioned previously[1] but such is the pace of work in this field that more recent experiments have already made it out of date to some extent. However, some general conclusions can safely be drawn.

The technique mainly used so far, both by Nirenberg and his colleague6 and by Ochoa and his group[7], has been to synthesize enzymatically "random" polymers of two or three of the four bases. For example, a polynucleotide, which I shall call poly (U,C), having about equal amounts of uracil and cytosine in (presumably) random order will increase the incorporation of the amino acids phenylalanine, serine, leucine, and proline, and possibly threonine. By using polymers of different composition and assuming a triplet code one can deduce limited information about the composition of certain triplets.

From such work it appears that, with minor reservations, each polynucleotide incorporates a characteristic set of amino acids. Moreover the four bases appear quite distinct in their effects. A comparison between the triplets tentatively deduced by these methods with the *changes* in amino acid sequence produced by mutation shows a fair measure of agreement. Moreover the incorporation requires the same components needed for protein synthesis, and is inhibited by the same inhibitors. Thus the system is most unlikely to be a complete artefact and is very probably closely related to genuine protein synthesis.

As to the actual triplets so far proposed it was first thought that possibly every triplet had to include uracil, but this was neither plausible on theoretical grounds nor supported by the actual experimental evidence. The first direct evidence that this was not so was obtained by my colleagues Bretscher and Grunberg-Manago[8], who showed that a poly (C,A) would stimulate the incorporation of several amino acids. Recently other worker[9,10] have reported further evidence of this sort for other polynucleotides not containing uracil. It now seems very likely that many of the 64 triplets, possibly most of them, may code one amino acid or another, and that in general several distinct triplets may code one amino acid. In particular a very elegant experiment II suggests that both (UUC) and (UUG) code leucine (the brackets imply that the order within the triplets is not yet known). This general idea is supported by several indirect lines of evidence which cannot be detailed here. Unfortunately it makes the unambiguous determination of triplets by these methods much more difficult than would be the case if there were only one triplet for each amino acid. Moreover, it is not possible by using polynucleotides of "random" sequence to determine the order of bases in a triplet. A start has been made to construct polynucleotides whose exact sequence is known at one end, but the results obtained so far are suggestive rather than conclusive[12]. It seems likely however from this and other unpublished evidence that the amino end of the polypeptide chain corresponds to the "right-hand" end of the polynucleotide chain - that is, the one with the 2', 3' hydroxyls on the sugar.

It seems virtually certain that a single chain of RNA can act as messenger RNA, since poly U is a single chain without secondary structure. If poly A is added to poly U, to form a double or triple helix, the combination is inactive. Moreover there is preliminary evidence[9] which suggests that secondary structure within a polynucleotide inhibits the power to stimulate protein synthesis.

It has yet to be shown by direct biochemical methods, as opposed to the indirect genetic evidence mentioned earlier, that the code is indeed a triplet code.

Attempts have been made from a study of the changes produced by mutation to obtain the relative order of the bases within various triplets, but my own view is that these are premature until there is more extensive and more reliable data on the composition of the triplets.

Evidence presented by several groups[8,9,11] suggest that poly U stimulates both the incorporation of phenylalanine and also a lesser amount of leucine. The meaning of this observation is unclear, but it raises the unfortunate possibility of ambiguous triplets; that is, triplets which may code more than one amino acid. However one would certainly expect such triplets to be in a minority.

It would seem likely, then, that most of the sixty-four possible triplets will be grouped into twenty groups. The balance of evidence both from the cell-free system and from the study of mutation, suggests that this does not occur at random, and that triplets coding the same amino acid may well be rather similar. This raises the main theoretical problem now outstanding. Can this grouping be

deduced from theoretical postulates? Unfortunately, it is not difficult to see how it might have arisen at an extremely early stage in evolution by random mutations, so that the particular code we have may perhaps be the result of a series of historical accidents. This point is of more than abstract interest. If the code does indeed have some logical foundation then it is legitimate to consider all the evidence, both good and bad, in any attempt to deduce it. The same is not true if the codons have no simple logical connection. In that case, it makes little sense to guess a codon. The important thing is to provide enough evidence to prove each codon independently. It is not yet clear what evidence can safely be accepted as establishing a codon. What is clear is that most of the experimental evidence so far presented falls short of proof in almost all cases.

In spite of the uncertainty of much of the experimental data there are certain codes which have been suggested in the past which we can now reject with some degree of confidence.

Comma-less triplet codes
All such codes are unlikely, not only because of the genetic evidence but also because of the detailed results from the cell-free system.

Two-letter or three-letter codes
For example a code in which A is equivalent to O, and G to U. As already stated, the results from the cell-free system rule out all such codes.

The combination triplet code
In this code all permutations of a given combination code the same amino acid. The experimental results can only be made to fit such a code by very special pleading.

Complementary codes
There are several classes of these. Consider a certain triplet in relation to the triplet which is complementary to it on the other chain of the double helix. The second triplet may be considered either as being read in the same direction as the first, or in the opposite direction. Thus if the first triplet is UCC, we consider it in relation to either AGG or (reading in the opposite direction) GGA.

It has been suggested that if a triplet stands for an amino acid its complement necessarily stands for the same amino acids, or, alternatively in another class of codes, that its complement will stand for no amino acid, i.e. be nonsense.

It has recently been shown by Ochoa's group that poly A stimulates the incorporation of lysine[10]. Thus presumably AAA codes lysine. However since UUU codes phenylalanine these facts rule out all the above codes. It is also found that poly (U,G) incorporates quite different amino acids from poly (A,C). Similarly poly (U,C) differs from poly (A,G)[9, 10]. Thus there is little chance that any of this class of theories will prove correct. Moreover they are all, in my opinion, unlikely for general theoretical reasons.

A start has already been made, using the same polynucleotides in cell-free systems from different species, to see if the code is the same in all organisms. Eventually it should be relatively easy to discover in this way if the code is universal, and, if not, how it differs from organism to organism. The preliminary results presented so far disclose no clear difference between E. coli and mammals, which is encouraging[10, 13].

At the present time, therefore, the genetic code appears to have the following general properties:

(1) Most if not all codons consist of three (adjacent) bases.
(2) Adjacent codons do not overlap.
(3) The message is read in the correct groups of three by starting at some fixed point.
(4) The code sequence in the gene is co-linear with the amino acid sequence, the polypeptide chain being synthesized sequentially from the amino end.
(5) In general more than one triplet codes each amino acid.
(6) It is not certain that some triplets may not code more than one amino acid, i.e. they may be ambiguous.
(7) Triplets which code for the same amino acid are probably rather similar.
(8) It is not known whether there is any general rule which groups such codons together, or whether the grouping is mainly the result of historical accident.
(9) The number of triplets which do not code an amino acid is probably small.
(10) Certain codes proposed earlier, such as comma-less codes, two- or three-letter codes, the combination code, and various transposable codes are all unlikely to be correct.

(11) The code in different organisms is probably similar. It may be the same in all organisms but this is not yet known.

Finally one should add that in spite of the great complexity of protein synthesis and in spite of the considerable technical difficulties in synthesizing polynucleotides with defined sequences it is not unreasonable to hope that all these points will be clarified in the near future, and that the genetic code will be completely established on a sound experimental basis within a few years.

The references have been kept to a minimum. A more complete set will be found in the first reference.

---

1. F.H.C. Crick in *Progress in Nucleic Acid Research*, J.N. Davidson and Waldo E. Cohn (Eds.), Academic Press Inc., New York (in the press).
2. H.G. Wittmann, *Z. Vererbungslehre*, 93 (1962) 491.
A. Tsugita, *J. Mol. Biol.*, 5 (1962) 284, 293.
3. F.H.C. Crick, L. Bamett, S. Brenner, and R.J. Watts-Tobin, *Nature*, 192 (1961) 1227.
4. M.A. Naughton and Howard M. Dintzis, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 1822.
5. G. von Ehrenstein and F. Lipmann, *Proc. Natl. Acad. Sci. U.S.*, 47 (1961) 941.
6. J.H. Matthaei and M.W. Nirenberg, *Proc. Natl. Acad. Sci. U.S.*, 47 (1961) 1580.
M.W. Nirenberg and J.H. Matthaei, *Proc. Natl. Acad. Sci. U.S.*, 47 (1961) 1588. M. W. Nirenberg, J. H. Matthaei, and O. W. Jones, P*roc. Natl. Acad. Sci. U.S.*, 48 (1962) 104.
J.H. Matthaei, O.W. Jones, R.G. Martin, and M.W. Nirenberg, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 666.
7. P. Lengyel, J.F. Speyer, and S. Ochoa, *Proc. Natl. Acad. Sci. U.S.*, 47 (1961) 1936.
J.F. Speyer, P. Lengyel, C. Basilio, and S. Ochoa, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 63.
P. Lengyel, J.F. Speyer, C. Basilio, and S. Ochoa, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 282.
J.F. Speyer, P. Lengyel, C. Basiho, and S. Ochoa, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 441.
C. Basilio, A.J. Wahba, P. Lengyel, J.F. Speyer, and S. Ochoa, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 613.
8. M.S. Bretscher and M. Grunberg-Manago, *Nature*, 195 (1962) 283.
9. O.W. Jones and M.W. Nirenberg, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 2115.
10. R.S. Gardner, A.J. Wahba, C. Basiho, R.S. Miller, P. Lengyel, and J.F. Speyer, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 2087.
11. B. Weisblum, S. Benzer, and R.W. Holley, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 1449.
12. A.J. Wahba, C. Basilio, J.F. Speyer, P. Lengyel, R.S. Miller, and S. Ochoa, *Proc. Natl. Acad.. Sci. U.S.*, 48 (1962) 1683.
13. H.R.V. Arnstein, R.A. Cox, and J.A. Hunt, *Nature*, 194 (1962) 1042.
E.S. Maxwell, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 1639.
I.B. Weinstein and A.N. Schechter, *Proc. Natl. Acad. Sci. U.S.*, 48 (1962) 1686.

From *Nobel Lectures, Physiology or Medicine 1942-1962*, Elsevier Publishing Company, Amsterdam, 1964