The Croonian Lecture, 1966

The genetic code

By F. H. C. Crick, F.R.S.

(*Delivered* 5 *May* 1966—*Received* 19 *October* 1966)

1. Introduction

(*a*) *The nature of the problem*

Genes are made of nucleic acid. Enzymes are made of protein. The amino acid sequence of a particular protein is synthesized under instruction from a particular piece of nucleic acid.

Each protein is made of one or more polypeptide chains, synthesized by condensing together amino acids, head to tail, with the elimination of water. A typical polypeptide chain is several hundred amino acid residues long. Nevertheless only twenty different kinds of amino acids are commonly found in proteins. This standard set of twenty is the same throughout nature.

Nucleic acid is made of polynucleotide chains. The repeating unit of the chain is a sugar (ribose for *RNA*, deoxyribose for *DNA*) connected to a phosphate. A base is joined on to each sugar. There are four common bases in nucleic acid. *DNA* usually has adenine, guanine, cytosine and thymine. In *RNA* thymine is replaced by uracil.

Thus protein is written in a twenty-letter language, nucleic acid in a four-letter language. *The genetic code* is the dictionary which connects the two languages. The group of bases which codes one amino acid is called a *codon*. It is now known that each codon consists of three adjacent bases. Notice that as far as we know the cell can translate in one direction only, from nucleic acid to protein, not from protein to nucleic acid. This hypothesis is known as the Central Dogma.

If one could compare the base sequence of a long piece of nucleic acid with the amino acid sequence for which it codes, the genetic code could easily be deduced. Unfortunately this direct approach is not yet possible because of the technical difficulties in determining a long nucleotide sequence. Consequently more indirect approaches must be used.

(*b*) *The biochemistry of protein synthesis*

This cannot be described here in detail, but some knowledge of it is necessary to understand part of the evidence for the genetic code. For a fuller account see for example the recent text-book by Watson (1965).

For most organisms the genetic material is double-stranded *DNA*. It does not function directly in protein synthesis. This role is played by single-stranded messenger *RNA*, which is a (complementary) copy of one of the strands of selected parts of the *DNA*. Messenger *RNA* is synthesized by a special enzyme, polynucleotide polymerase, which needs *DNA* as a template.

The actual site of protein synthesis is the ribosome. Ribosomes are complex structures, about 200 Å in diameter, made of *RNA* and protein in (very roughly) equal proportions. Each ribosome consists of two parts, one approximately twice the size of the other. The ribosome moves along the messenger *RNA*, 'reading' the message in the base sequence and synthesizing a polypeptide chain one step at a time, starting from the amino end.

Each amino acid is activated by its own special enzyme, which uses one molecule of *ATP* to synthesize a mixed anhydride of the amino acid and *AMP*. This compound is tightly bound to the enzyme. The enzyme then transfers the amino acid to the terminal ribose of special *RNA* molecules (each about 80 bases long) known as transfer *RNA* (*tRNA*) or sometimes as soluble *RNA* (*sRNA*). There are one or a small number of types of *tRNA* for each amino acid.

The *tRNA* carries the amino acid to the ribosome, and is responsible for recognizing the next codon on the messenger *RNA* (*mRNA*). It probably does this by forming base-pairs between the three bases of the codon (on the messenger *RNA*) and three of its own bases, known as its *anticodon*.

At any moment in the middle of the synthesis of a polypeptide chain the partly formed chain is joined by its carboxyl end to the molecule of *tRNA* which inserted the last amino acid. Adjacent to it on the ribosome comes the *tRNA* for the next amino acid indicated by the *mRNA*. The basic step of protein synthesis is to transfer the polypeptide chain from the first *tRNA* to the amino acid joined to the second *tRNA*, thus lengthening the chain by one residue. The first *tRNA* then returns to solution, and the mechanism resets in some way in preparation for the next step.

Protein synthesis can take place in the test-tube using the parts of the system described above, together with *GTP* and several soluble factors. Molecules of single-stranded *RNA*, added to the system, can often act as messenger *RNA* and direct the synthesis of polypeptide chains.

(c) *Early work on the genetic code*

The work up to 1962 has been briefly reviewed by Crick (1963*a*). The more detailed review (Crick 1963*b*) should be consulted for references. This work suggested that the code has the following features: (i) each codon consists of three (consecutive) bases; (ii) adjacent codons do not overlap. Thus any base in the messenger *RNA* belongs to only one codon; (iii) most of the sixty-four possible codons represent one amino acid or another, i.e. the code is 'degenerate'; (iv) triplets which code the same amino acid are often rather similar; (v) the code is probably universal, or nearly so, meaning that it is largely the same in all organisms.

The evidence that the code was not overlapping was mainly due to the study of the changes in amino acid sequence produced by mutation. This was almost always a change to a single amino acid in the sequence, as would be expected for a non-overlapping code.

The evidence that the code was a triplet code came from the study of phase-shift mutants in the $r_{II}B$ cistron of phage $T4$. Whereas the addition of one or two bases to a gene put the reading out of phase, the adding (by genetic means) of three bases put the message back into phase again.

The probable composition (but not the base sequence) of many of the sixty-four codons was suggested by work on the cell-free system for protein synthesis, using added messenger *RNA* of known composition but random sequence made by the enzyme polynucleotide phosphorylase. This (and the genetic evidence) made it likely that most triplets stood for one amino acid or another and that triplets representing the same amino acid were probably somewhat alike, as hinted by the limited types of amino acid change found in mutants.

(*d*) *Collinearity of gene and protein*

It was long suspected that a gene and the protein it codes are collinear. In other words, that the order of codons along the gene is the same as the order of the corresponding amino acids along the polypeptide chain of the protein. This was first proved by Yanofsky and his colleagues (Yanofsky *et al.* 1964) for the *A* protein of the enzyme tryptophan synthetase of *E. coli.* They determined the amino acid sequence of a stretch of polypeptide chain, about seventy-five amino acids long, and located on it the separate amino acid alterations produced by nine different mutations. They also found, by purely genetic methods, the order of these mutations on the genetic map. The results showed that the two orders were the same. Moreover sites close together on the polypeptide chain were also close together on the genetic map.

The same result was also obtained by Sarabhai, Stretton, Brenner & Bolle (1964), by a neat method, using chain-terminating mutants of the gene for the head protein of bacteriophage *T*4.

It has not yet been shown directly that the actual *DNA* and the protein it codes are collinear. However, recently Hogness (1966) has produced evidence for phage λ that the order of five genes on the genetic map is the same as their order on the *DNA* of λ.

(*e*) *The direction of reading*

Amino acid sequences are conventionally written starting at the amino end of the polypeptide chain. Nucleic acid sequences are written with the 5′ hydroxyl on the left. It is a matter for experiment to decide whether the relationship *between* the *mRNA* sequence and the amino acid sequence for which it codes is the same as implied by these conventions, or the reverse. Notice this is not precisely the same question as the *time-sequence* of reading the message.

It has fortunately turned out that, in spite of earlier results to the contrary, the two conventions agree. That is, the 5′ end of the *mRNA* codes for the amino end of the polypeptide chain. The lines of evidence supporting this are as follows:

(i) From polymers of known sequence. For example it has been shown that a messenger *RNA* of the form AAAA...AAAAC makes a polypeptide consisting mainly of lysine (AAA) but with some asparagine (AAC) at the C-terminal end (Salas *et al.* 1965). Studies on polymers which have special sequences at the 5′ end of the *mRNA* also support this direction of reading (Smith, Salas, Stanley, Wahba & Ochoa 1966; Stanley, Salas, Wahba & Ochoa 1966).

(ii) From the deciphering of the message using amino acid sequences from double frame-shift mutants. This is explained in §6.

(iii) By a combination of genetic and biochemical methods, using the genetic code. This has been done by Yanofsky using the *A* protein of tryptophan synthetase (Guest & Yanofsky 1966).

2. The genetic code to-day

The genetic code is now known with reasonable certainty, at least for *Escherichia coli*. It is customary to display it in the compact form shown in figure 1. Each entry in the figure corresponds to a particular triplet, and shows, using the standard abbreviations, which amino acid is coded by that triplet, Thus the entry indicated by CAU is labelled 'His' implying that the triplet CAU (cytosine-adenine-uracil) codes histidine.

2nd → 1st ↓	U	C	A	G	3rd ↓
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	ochre	?	A
	Leu	Ser	amber	Tryp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	GluN	Arg	A
	Leu	Pro	GluN	Arg	G
A	Ileu	Thr	AspN	Ser	U
	Ileu	Thr	AspN	Ser	C
	Ileu	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure 1. The four standard bases, uracil, cytosine, adenine and guanine are represented by the letters U, C, A and G respectively. The first base of any triplet is indicated on the left, the second base at the top and the third at the right of the figure. The twenty amino acids are represented by their standard abbreviations. Thus 'Phe' stands for phenylalanine, etc. The triplets marked 'ochre' and 'amber' are believed to signal the termination of the polypeptide chain. Those associated with chain initiation are not marked in this figure.

In what follows I have displayed some of the experimental evidence which supports figure 1, in order to show both the methods used and the strength of the experimental support. This summary is not meant to be comprehensive. In any case new evidence is accumulating so rapidly that any summary would soon be out of date. The evidence falls roughly into two main classes: (i) data obtained using the cell-free system for protein synthesis—these techniques allocate triplets to particular amino acids (e.g. the binding test: repeating *mRNA*s); (ii) data obtained from synthesis in intact cells—these techniques show the relationship *between* triplets (e.g. mutagenic changes: double phase-shift results).

In addition we have some indication of the base sequence of several anticodons on the *tRNA*.

For a more extensive account of these and other related topics the reader is referred to volume XXXI of the Cold Spring Harbor Symposia on Quantitative Biology entitled *The genetic code*. The meeting at which these papers were presented took place early in June 1966. The volume is not only up to date but contains more detailed references than it has been possible to give here.

3. The triplet binding test

This was initiated by Nirenberg & Leder (1964) who devised a new technique for allocating a codon to an amino acid. They showed that in the presence of a triplet (a trinucleoside diphosphate) the ribosomes would preferentially bind the corresponding species of *tRNA*.

The mixture of triplet, ribosomes and *tRNA* molecules (which had been labelled with a single radioactive amino acid) was allowed to stand together, in the presence of a suitable Mg^{2+} concentration, to permit binding to take place. It was then passed through a Millipore filter, which retained both the ribosomes and the bound *tRNA*, while allowing the unbound *tRNA* to pass through. The radioactivity on the filter was then counted. The control was to repeat the identical experiment without the triplet.

This technique has been used extensively by Nirenberg and his colleagues (Leder & Nirenberg 1964*a*, *b*; Bernfield & Nirenberg 1965; Trupin *et al.* 1965; Nirenberg *et al.* 1965; Brimacombe *et al.* 1965) and also by Khorana and his colleagues (Söll *et al.* 1965; Söll *et al.* 1966). All sixty-four triplets have been synthesized, by one method or another, and each has been tested against many of the twenty amino acids. In the majority of cases the binding for one particular *tRNA* has been several times the background in the control.

Unfortunately the method cannot always be completely trusted since it sometimes gives practically no binding in cases where, from other experiments, a positive response would be expected; conversely, in some cases the *tRNA*s belonging to several amino acids respond to one triplet. It is suspected that most of these weaker responses are artifacts of the method.

However, when the binding test gives a strong, unambiguous result it is probably safe to accept it. In figure 2 I have set out all these results (published or in press) which have given a binding at least three times the background. This criterion is rather arbitrary, as a more precise assessment would take into account the exact details of the binding test used in each case (for example, whether purified *tRNA* was used) and the absolute size of the binding. Moreover the binding varies somewhat from experiment to experiment. Almost all the results are for *E. coli*, but AGA and AGG for arginine have only been convincingly shown for yeast.

2nd → 1st ↓	U	C	A	G	3rd ↓
U	Phe	Ser	Tyr	Cys(Val)	U
	Phe	(Ser)	Tyr	Cys	C
	—	Ser	—	—	A
	Leu	Ser	—	—	G
C	—	—	His	Arg	U
	—	Pro	His	Arg	C
	—	Pro	GluN	Arg	A
	Leu	—	GluN	—	G
A	Ileu	Thr	AspN	—	U
	Ileu	Thr	AspN	—	C
	—	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	—	Gly	G

Figure 2. Conventions as for figure 1. The entries show the triplets which have given a response of greater than three times the background in the binding test. Thus most of the allocations can be considered as probable.

It can be seen that just over three-quarters of the triplets can be allocated in this way. There is one result which is almost certainly false—namely the response of valine to UGU. Many of the weak additional bindings (not shown in figure 2) occur as a response to XAB, when the expected triplet is ABY. This often happens when the binding to ABY is strong, and is presumably due to a small response to the AB doublet, read out of phase.

A modified binding test has been used by Matthaei *et al.* (1966), using polymers of the form XpYpZ...pZ, about 30 residues long. They claim that this often shows the binding expected of XYZ preferentially (apart from the binding expected due to ZZZ). The method had earlier given results believed to be false, and in any case the binding is usually rather small compared to the background. In spite of claims to the contrary no triplets can be confidently allocated by this method which have not already been allocated by the triplet-binding method.

4. *mRNA* WITH A DEFINED BASE SEQUENCE

(*a*) *Polynucleotides with repeating sequences*

This approach has been pioneered by Khorana (Nishimura, Jones, Ohtsuka, Hayatsu, Jacob & Khorana 1965; Nishimura, Jones & Khorana 1965). In outline their method had been to synthesize chemically a small repeating sequence, about 10 or 12 residues long, of double-stranded complementary *DNA*, each strand being synthesized separately. This has then been given to the enzyme *DNA* polymerase as a template. The product is double-stranded *DNA*, with the same repeating sequence but a much longer chain length. This *DNA* was then used as a template for the enzyme *RNA* polymerase. By supplying for the synthesis only some, but not all, of the four nucleoside triphosphates it was possible to copy either one of the two *DNA* strands, without at the same time copying the other. The product was a long *RNA* molecule which was shown, by nearest-neighbour analysis, to have the expected repeating sequence to a high degree of accuracy. These *RNA* molecules were then used as messenger *RNA* in the cell-free system for protein synthesis.

TABLE 1.

type of *mRNA*	expected polypeptides
$(AB)_n$	$\ldots\alpha\beta\alpha\beta\alpha\beta\ldots$
$(ABC)_n$	$\left.\begin{array}{r}\ldots\alpha\alpha\alpha\alpha\alpha\ldots\\ \text{and }\ldots\beta\beta\beta\beta\beta\ldots\\ \text{and }\ldots\gamma\gamma\gamma\gamma\gamma\ldots\end{array}\right\}$
$(ABCD)_n$	$\ldots\alpha\beta\gamma\delta\alpha\beta\gamma\delta\ldots$

TABLE 2.

mRNA	polypeptide(s) produced
$(UC)_n$	$(Ser.Leu)_m$
$(UG)_n$	$(Cys.Val)_m$
$(AG)_n$	$(Arg.Glu)_m$
$(AC)_n$	$(Thr.His)_m$
$(AAG)_n$	$(Lys)_m+(Arg)_m+(Glu)_m$
$(UUG)_n$	$(Leu)_m+(Cys)_m+(Val)_m$
$(GUA)_n$	$(Val)_m+(Ser)_m$
$(UAC)_n$	$(Tyr)_m+(Thr)_m+(Leu)_m$
$(AUC)_n$	$(Ileu)_m+(Ser)_m+(His)_m$

If the code is a triplet code we expect the results shown in table 1. The fact that this type of result is obtained establishes by direct biochemical methods that the code is a triplet code. Khorana's results to date (Khorana, personal communication) are summarized in table 2.

The fact that $(UC)_n$ codes for the repeating polypeptide ...Leu.Ser... does not establish by itself the amino acid to be associated with the two triplets involved, namely UCU and CUC. However, since the binding test shows that UCU codes serine, we may deduce with confidence that CUC codes leucine.

(*b*) *Polynucleotides with non-repeating sequences*

The synthesis of these, except in special cases, is usually rather laborious, and only a limited number of different *mRNA* molecules have been made so far. Some examples are given in Thach, Dewey, Brown & Doty (1966) and Stanley *et al.* (1966). The latter paper strongly suggests that AUA codes isoleucine.

5. Amino acid changes in mutants

The typical amino acid change is to a single amino acid in the entire polypeptide chain. It may occur 'spontaneously' or it may be produced by the action of a mutagen. In some case the changes to the base sequence produced by a particular mutagen may be strongly suspected on chemical grounds. Thus hydroxylamine is believed to attack only cytosine, and not the other three bases.

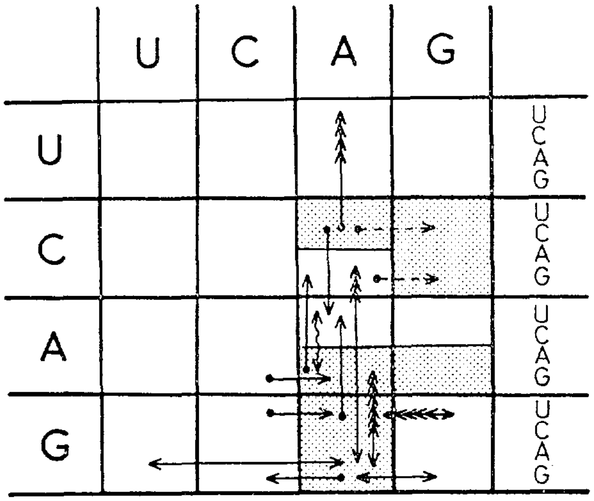

Figure 3. The figure shows the genetic code, but for convenience the names of the amino acids have been omitted. The arrows indicate the amino acid changes for thirty-six distinct abnormal human haemoglobins, each arrowhead representing one example. The base change can be deduced in all cases except that shown by a wavy arrow and the two dotted arrows. The stippled areas represent charged amino acids.

In most cases mutants are screened or selected by one method or another. For example most abnormal human haemoglobins have been picked up because they are electrophoretically different from normal adult haemoglobin. Consequently the amino acid change is likely to involve a change of charge. In other cases there may be a bias towards mutants which destroy the action of the gene, and thus 'nonsense' mutants may be preferentially selected.

It is convenient to display the change of an amino acid, due to a mutation, as an arrow on the figure of the genetic code (see figure 1). We now know that almost all such changes are due to an alteration of a single base in the genetic nucleic acid.

If the change is to a base in the first position of a codon, the arrow marking the change will be vertical, and will begin and end in the same relative position within a square on the figure. If the change is to the second base in the codon the arrow will be horizontal. If to the third base the arrow will be vertical, but will begin and end in the same square.

Thus in all such cases the arrow will be either horizontal or vertical. A diagonal arrow would imply that at least two bases in the codon had been changed. If

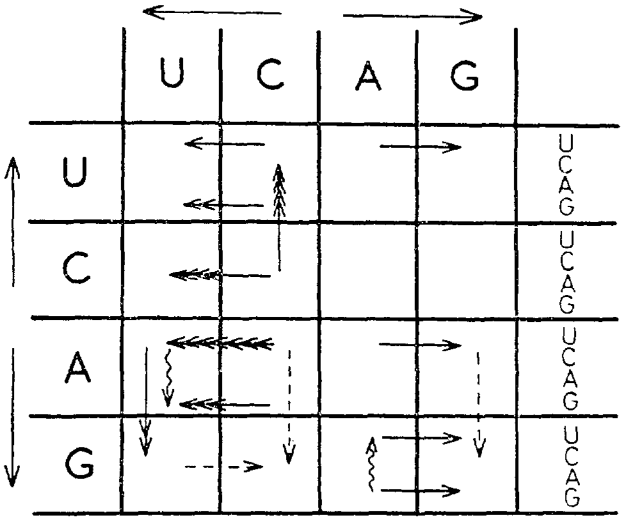

FIGURE 4. The figure shows the genetic code, but for convenience the names of the amino acids have been omitted. The arrows indicate the amino acid changes found by Wittmann among mutants of tobacco mosaic virus. The mutagen used was nitrous acid except for the three cases shown by dotted arrows when fluorouracil was employed. The wavy arrows imply that the base-changes in those cases are uncertain. The arrows outside the table show the changes expected from the action of nitrous acid.

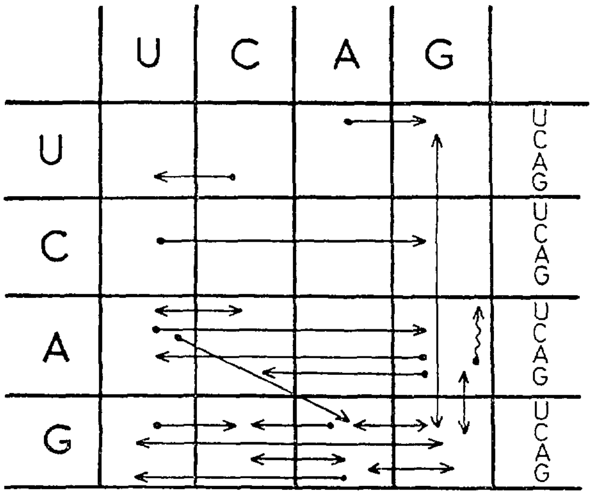

FIGURE 5. The figure shows the genetic code, but for convenience the names of the amino acids have been omitted. The arrows indicate the amino acid changes found by Yanofsky in the A protein of the tryptophan synthetase of *Escherichia coli*. Although most of these changes have been found many times only a single arrow is shown here for each change. The wavy arrow implies that the base-change in this case is uncertain.

changes in amino acids were arbitrarily made at random, roughly half the arrows would be horizontal or vertical and half diagonal.

The results for the three largest sets of mutagenic changes are set out in figures 3–5. It can be seen that with the exception of one arrow in figure 5 all the arrows are either vertical or horizontal. This is striking confirmation of the genetic code.

Although the code is degenerate it is often possible to deduce the actual change in bases involved in a mutation, even if the precise codons involved cannot be determined. For example, the change phe → tyr must go from either UUU or UUC to UAU or UAC respectively. In either case the mutation must have been a change from U to A. It is always possible to deduce the base which has changed unless (i) leucine, serine or arginine are involved or (ii) the change is within a square in the figure, i.e. involves the third base of the triplet. Even in some of these cases it may be possible to deduce the base-change unambiguously. Thus most amino acid changes imply a particular base change.

Changes of a pyrimidine to a pyrimidine (U → C, or C → U) and of a purine to a purine (A → G, or G → A) are called 'transitions'. Changes from a purine to a pyrimidine, or vice versa are called 'transversions'.

Figure 3 shows the results for thirty-six different abnormal human haemoglobins. The data are taken from table 10·1 of Lehmann & Huntsman (1966), which should be consulted for references. Those areas of the figure representing charged amino acids have been stippled. It can be seen that, as expected from the method of detection of the mutants each arrow either starts or finishes in a stippled region. Notice that some changes occurred in both directions—the corresponding arrows are double-headed. Both transitions and transversion occur, but the former are rather more common.

Figure 4 shows most of Wittmann's results for tobacco mosaic virus (see the review by Wittmann & Wittmann-Liebold 1966; similar results have also been reported by Tsugita 1962). The genetic material of this virus is single-stranded *RNA*, which is believed to act also as the messenger *RNA* for the virus. All but three of these mutants were produced when nitrous acid was used as a mutagen, either on the whole virus or on its *RNA*. Nitrous acid changes C to U. It changes A to hypoxanthine, which is then copied as if it were a G, so the effective change is A to G. The change of G to xanthine is believed to be lethal. The figure shows that each arrow points in a single direction only, as would be expected, and all but one could fit the base changes expected. This exceptional change (Glu → Asp) may have been due to an unexpected action of nitrous acid, or more likely to a spontaneous mutant which was accidentally picked up with the nitrous acid mutants.

The other three mutations in figure 4 were due to fluorouracil added during viral growth. This is expected to produce the changes A → G or U → C. All three cases fit this expectation.

Figure 5 shows the change picked up by Yanofsky and his co-workers as mutations, or back-mutations, of the *A* protein of tryptophan synthetase of *E. coli* (see the review by Yanofsky 1966). In this case no particular base-change is favoured although detailed studies by Yanofsky show that the base changes produced by various mutagens are almost always those expected.

Here we have one diagonal arrow, due to the change Ileu → Asp. This may have been due to two bases having been altered. An alternative explanation is that the actual change was Ileu → AspN, and that the asparagine was then deaminated by some unknown process to give aspartic acid. Yanofsky (1966) has presented experimental evidence which makes this seem probable.

There is one other change reported recently by Yanofsky (1966, and personal communication) which has not been incorporated in figure 5. This is the change, at position 48 in the protein, of glutamine to methionine, This implies GAG (or GAA) to AUG, and thus a change of at least two bases. The interesting point about this mutation is that it has been picked up only once, and has never back-mutated to wild. These two facts in themselves would make one suspect that it is the rare case of a change of two bases, and for this reason it has not been included in figure 5.

6. Amino acid changes produced by phase shifts

Since the messenger *RNA* is read sequentially three bases at a time the addition of a nucleotide at one point will throw the whole of the subsequent reading out of phase. The reading can be brought back into phase again, however, if a nucleotide is subtracted at some subsequent point, although the message will be misread in between the two alterations. This effect was predicted by Crick, Barnett, Brenner & Watts-Tobin (1961) from genetic studies. It has recently been confirmed directly by Streisinger and his colleagues (Terzaghi *et al.* 1966).

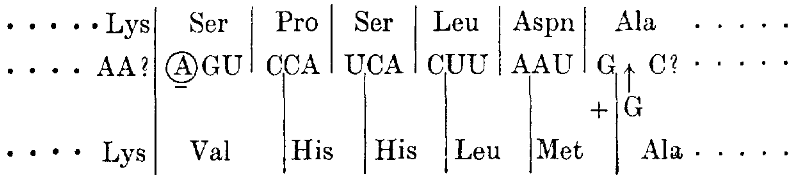

Figure 6. The amino acid sequences shown are for part of the protein lysozyme from phage *T*4 studied by Streisinger and his co-workers. The top line represents the wild-type sequence. The bottom line that found in the double phase-shift mutant. The base sequence in the middle has been deduced from the genetic code. It suggests that the first mutant lacked an 'A' and that the second had added a 'G' to the wild-type sequence.

The protein studied was the lysozyme of coliphage *T*4. Mutants were produced by means of acridines, which are strongly suspected to produce additions or subtractions of a base (or bases) rather than the alteration of one base into another. Such mutants typically destroy the function of a gene. By genetic methods a pair of such mutants were combined in one gene. This gene produced an altered protein having some enzymic activity.

The change produced in the amino acid sequence was found to be due to five adjacent amino acids. From the wild-type sequence, and the altered sequence in the double mutant, it was possible, using the genetic code, to deduce the probable base sequence of the messenger *RNA*. This is shown in figure 6. It will be seen that the first mutation was probably the deletion of an A, and the second the addition of a G.

It was found by trial that no other solution was compatible with the code and that no solution at all could be found if the direction of reading was considered to be in the reverse direction.

The data shown in figure 6 are useful in that they confirm certain doubtful allocations, such as UUA for leucine. Further examples of phase-shift changes in this region of lysozyme are given in the previous reference and the review by Streisinger *et al.* (1966).

7. The anticodon on the *tRNA*

Some progress has been made in locating the anticodon on several of the *tRNA* molecules. It was strongly suspected that the first two bases of the codon would pair with the corresponding bases on the anticodon using the standard base-pairs found in *DNA*. The relevant pairs for *RNA* are

A—U and G—C.

When the sequence of the alanine *tRNA* from yeast was worked out by Holley and his collaborators (Holley *et al.* 1965) they pointed out the triplet IGC as a

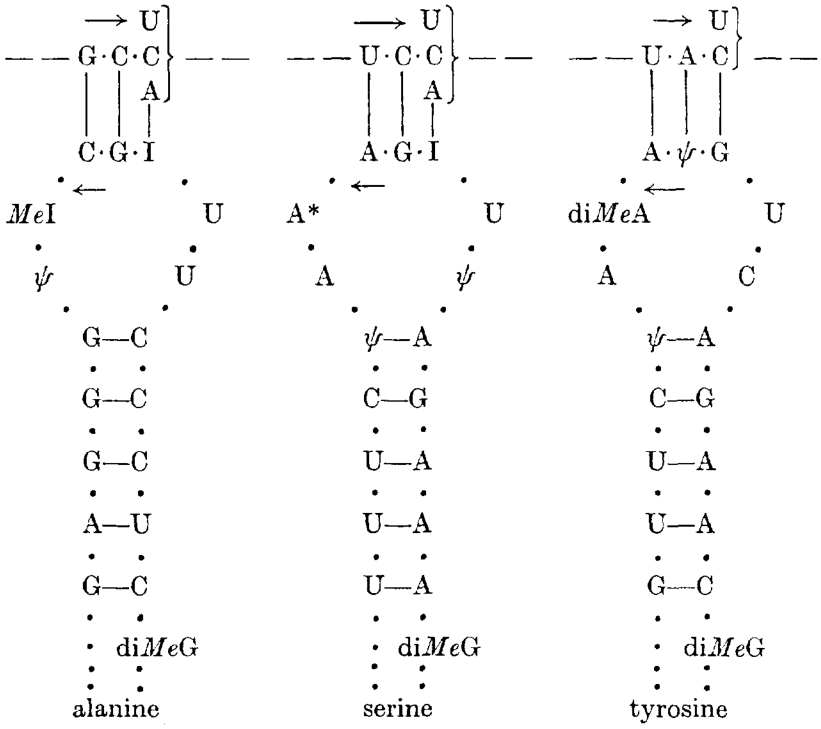

Figure 7. To show the anticodons in the base-sequence of the alanine, serine and tyrosine *tRNA*s from yeast (for references, see text). The anticodons occur near the middle of each *tRNA*. Only a part of each sequence is shown here. The dots indicate the run of the chain. The lines symbolize base-pairing. The codons on the messenger *RNA*, predicted by wobble theory, are shown at the top. The arrows show the directions of the polynucleotide chain, running from the 5′ to the 3′ end.

probable anticodon. This could pair in an *anti*-parallel manner with triplets of the form GCX (X = unknown) which code alanine. The secondary structure of *tRNA* is not known but a reasonable guess was that IGC occurred in the middle of a loop of seven bases linking the two ends of a short double-helical region having five base-pairs. The unusual base dimethylguanine (*diMeG*) occurred next to the start of this helix.

More recently Zachau and his colleagues (Dütting, Karau, Melchers & Zachau 1965) have presented the sequence of two closely related *tRNA*s for serine, and Madison (Madison, Everett & Kung 1966) that for a tyrosine *tRNA*, all from yeast. In all cases a similar structure can be drawn, containing the expected anticodon, as shown in figure 7. The only surprise is that pseudouridylic acid occurs in the tyrosine anticodon instead of uracil, but as these two bases can form the relevant hydrogen bonds with the A on the codon in the same way, this is not a violation of the pairing rules.

Table 3. Predicted rules for pairing in the third place of the codon.

anticodon	codon
U	A G
C	G
A	U
G	U C
I	U C A

The pairing of the third base of the codon is less clear. It now seems likely that one *tRNA* molecule can recognize several related codons, which only differ in the third place in the codon. Considerations of the base-pairs to be plausibly expected if there were some wobble in the recognition at this position has led me (Crick 1966) to the scheme for base-pairing in the third position of the codon shown in table 3. The present limited experimental evidence tends to support this scheme. (Söll *et al.* 1966; Kellogg, Doctor, Loebel & Nirenberg 1966). If enough instances accumulate to establish the recognition patterns shown in table 3 it should be possible to confirm codons by discovering the anti-codons on the various *tRNA* molecules. In any case it seems highly likely that the first two bases of any codon can be confirmed in this way, provided the relevant *tRNA* is not there in such small quantities as to make it technically difficult to prepare it in pure form.

8. Punctuation marks

(*a*) *Chain initiation*

This has only recently been discovered. In *E. coli* there is a special *tRNA* for methionine which is involved in the initiation of the polypeptide chain (Marcker & Sanger 1964). The methionine on this special *tRNA* has its amino group formylated by a special enzyme, which uses formyl-tetrahydrofolic acid as a formyl donor

(Marcker 1965). This *tRNA* appears to recognize the codons AUG and GUG (and possibly UUG and CUG) (Clark & Marcker 1966*a*; Kellogg *et al.* 1966). It is suspected that many if not all polypeptide chains in *E. coli* start with formylmethionine when first synthesized, but that some part of the beginning of the chain is then removed by a special enzyme or enzymes (Adams & Capecchi 1966; Webster, Engelhardt & Zinder 1966; Capecchi 1966). It is a striking fact, first discovered by Waller (1963), that 40 % of the amino ends of the polypeptide chains in *E. coli* start with methionine.

Even when the formyl group is not present, the special *tRNA* inserts methionine only at the beginning of the polypeptide chain, not in the middle (Clark & Marcker 1965, 1966*a*). Thus the property of initiation, which probably means that the *tRNA* enters directly the slot on the ribosome normally occupied by polypeptide-*tRNA* (Bretscher & Marcker 1966), is due to the special nature of this *tRNA*, not merely to the formyl group. However the formyl group probably accelerates the initiation (Clark & Marcker 1966*b*). There is another more normal *tRNA* for methionine which cannot be formylated. This *tRNA* is not concerned with chain initiation but puts methionine into the middle of polypeptide chains. It responds only to the codon AUG (Clark & Marcker 1966*a*).

Note that the triplet GUG appears to stand for valine in the middle of a chain but for methionine at the beginning. This is not surprising when it is realized that to begin a chain the initial *tRNA* probably has to read the messenger *RNA* in a different slot on the ribosome from the one usually used for first recognition between a *tRNA* molecule and the messenger *RNA*.

It has been shown in the *in vitro* system that the triplet AUG acts as a 'phaser' when it is near the beginning of an *RNA* chain, putting the reading mechanism into the phase defined by the position of the AUG (Sundararajan & Thach 1966; Thach *et al.* 1966). The triplet GUG has not yet been tested in this way. It is not yet known whether there are other methods of chain initiation in *E. coli*, nor what the mechanism is in higher organisms.

(*b*) *Chain termination*

Some positive step is probably necessary for chain germination, since during protein synthesis the growing polypeptide chain is always joined to the *tRNA* molecule brought to the ribosome with the last amino acid to be incorporated (Gilbert 1963). To produce a free peptide this link must be severed after the last amino acid of the chain has been added.

The two triplets UAA (ochre) and UAG (amber) are believed to produce termination of the polypeptide chain (Weigert & Garen 1965; Brenner, Stretton & Kaplan 1965). The trivial names 'ochre' and 'amber' refer to sets of mutations in many different genes which have been characterized by their suppression properties. Special suppressor genes exist which partly suppress the reading of UAG, or of UAA *and* UAG, but not of UAA alone. These suppressors act by occasionally putting in an amino acid instead of always terminating the chain. The amino acid inserted is characteristic of the suppressor. Thus several different suppressors of UAG are known; one puts in tyrosine, another glutamine and a third serine, (for references see table 2 of Kaplan, Stretton & Brenner 1965). Such suppression is probably due in each case to a genetic alteration of a particular *tRNA* (Capecchi & Gussin 1965; Smith, Abelson, Clark, Goodman & Brenner 1966). It has not yet been proved that the alteration is to the anticodon of the *tRNA*.

All the above applies to *mutations* which produce premature chain termination. The mechanism of *natural* chain termination is not yet known. It is suspected that the codon UAA is mainly involved, and that there is a special *tRNA* for chain termination. This has yet to be discovered.

Studies using the *in vitro* system for protein synthesis have shown that chain release occurs when either of the random polynucleotides poly UA or poly UAI is used as messenger *RNA* (Bretscher, Goodman, Menninger & Smith 1965; Takanami & Yan 1965; Ganoza & Nakamoto 1966) as would be expected from the composition of the triplets UAA and UAG.

9. Misreading and ambiguity

Partial misreading may be of several kinds. It may be produced by antibiotics, such as streptomycin, or by genetic defects in parts of the reading mechanism, as in extragenic suppression. Misreading may also occur in the *in vitro* system, for example if the Mg^{2+} concentration is too high, or the temperature too low. These topics will not be discussed in detail here.

A more serious problem is whether in a *normal* cell a triplet can be read in more than one way. This is known as 'ambiguity'. There is unfortunately some evidence that ambiguity may occur (von Ehrenstein 1966; Rifkin, Hirsch, Rifkin & Konigsberg 1966) though this is not yet clearly established. However, it is highly probable that only a few triplets will be ambiguous and that most of them can be read in only one way.

10 Universality

The twenty standard amino acids are the same throughout nature, as are the standard four bases. It does not follow, however, that the genetic code relating them is always the same.

Earlier evidence, not detailed here, using artificial messengers in the *in vitro* system had not uncovered any obvious differences in the genetic code between species, nor had experiments using mixed *in vitro* systems, in which the *mRNA* and the ribosomes came from one species and the *tRNA* from another. The best evidence to date is probably the excellent agreement between the code deduced for *E. coli* and the mutagenic data detailed in §5 derived from tobacco plants or human beings. There is thus little doubt that the genetic code is similar in most organisms. Whether there are any organisms which use a slightly modified version of the code remains to be seen. It is certainly possible that the initiation triplets may differ in different species.

11. Conclusion

From what has been said it can be seen that the genetic code is already known in outline. Further work is required to check the details, especially those concerning the punctuation marks, and to extend the results to other species. It seems very unlikely that the results shown in figure 1 will need radical alteration.

Much work still remains to be done on the exact biochemical mechanism of protein synthesis, a subject hardly touched on in this review. Beyond that the major unsolved problem is that of control mechanisms. In particular we do not yet know the base sequences which signal the beginning and end of a gene or an operon, nor how they are related, if at all, to the genetic code proper. Nor do we understand much about the control of the *rate* at which genes act.

Concerning the genetic code itself the structure of the code represents a problem of a very different kind. Has it some stereochemical basis or is it mainly the result of historic accident? These questions lead us back to the origin of life, a fascinating but difficult field in which speculation is rampant and really pertinent facts are hard to come by. This topic is outside the scope of this review.

The importance of the detailed work on the genetic code described here is not merely that it has uncovered the most important and central biochemical mechanisms in biology. The very existence of this exact knowledge establishes the general theoretical framework which has guided investigators for the last dozen years, and shows clearly the very different roles played in living things by nucleic acid and by protein. It demonstrates clearly how natural selection can operate at the molecular level, and illuminates such concepts as the absence of the inheritance of acquired characteristics. We can now confidently look forward to placing increasing areas of biology on a molecular basis.

References

Adams, J. M. & Capecchi, M. R. 1966 *Proc. Nat. Acad. Sci. U.S.* **55**, 147.

Bernfield, M. R. & Nirenberg, M. W. 1965 *Science* **147**, 479.

Brenner, S., Stretton, A. O. W. & Kaplan, S. 1965 *Nature, Lond.* **206**, 944.

Bretscher, M. S., Goodman, H. M., Menninger, J. R. & Smith, J. D. 1965 *J. Mol. Biol.* **14**, 634.

Bretscher, M. S. & Marcker, K. A. 1966 *Nature, Lond.* **211**, 380.

Brimacombe, R., Trupin, J., Nirenberg, M., Leder, P., Bernfield, M. & Jaouni, T. 1965 *Proc. Nat. Acad. Sci. U.S.* **54**, 954.

Capecchi, M. R. & Gussin, G. 1965 *Science* **149**, 417.

Capecchi, M. R. 1966 *Proc. Nat. Acad. Sci. U.S.* **55**, 1517.

Clark, B. F. C. & Marcker, K. A. 1965 *Nature, Lond.* **207**, 1038.

Clark, B. F. C. & Marcker, K. A. 1966*a* *J. Mol. Biol.* **17**, 394.

Clark, B. F. C. & Marcker, K. A. 1966*b* *Nature, Lond.* **211**, 378.

Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. 1961 *Nature, Lond.* **192**, 1227.

Crick, F. H. C. 1963*a* *Science*, **139**, 461.

Crick, F. H. C. 1963*b* In *Progress in nucleic acid research*, vol. 1, New York: Academic Press, Inc.

Crick, F. H. C. 1966 *J. Mol. Biol.* **19**, 548.

Dütting, D., Karau, W., Melchers, F. & Zachau, H. F. 1965 *Biochem. Biophys. Acta* **108**, 194.

Ganoza, M. C. & Nakamoto, T. 1966 *Proc. Nat. Acad. Sci. U.S.* **55**, 162.

Gilbert, W. 1963 *J. Mol. Biol.* **6**, 389.

Guest, J. R. & Yanofsky, C. 1966 *Nature, Lond.* **210**, 799.

Hogness, D. 1966 *Science* **153**, 94.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R. & Zamir, A. 1965 *Science*, **147**, 1462.

Kaplan, S., Stretton, A. O. W. & Brenner, S. 1965 *J. Mol. Biol.* **14**, 528.

Kellogg, D. A., Doctor, B. P., Loebel, J. F. & Nirenberg, M. W. 1966 *Proc. Nat. Acad. Sci. U.S.* **55**, 912.

Leder, P. & Nirenberg, M. W. 1964*a* *Proc. Nat. Acad. Sci. U.S.* **52**, 420.

Leder, P. & Nirenberg, M. W. 1964*b* *Proc. Nat. Acad. Sci. U.S.* **52**, 1521.

Lehmann, H. & Huntsman, R. G. 1966 *Man's haemoglobin*, Amsterdam: North-Holland Publishing Co.

Madison, J. T., Everett, G. A. & Kung, H. 1966 *Science*, **153**, 531.

Marcker, K. & Sanger, F. 1964 *J. Mol. Biol.* **8**, 835.

Marcker, K. A. 1965 *J. Mol. Biol.* **14**, 63.

Matthaei, H., Heller, G., Voigt, H-P., Neth, R., Schöch, G. & Kübler, H. 1966 In *Proc. 3rd Meeting European Biochem. Socs. Warsaw.* New York: Academic Press.

Nirenberg, M. & Leder, P. 1964 *Science*, **145**, 1399.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. & O'Neale, C. 1965 *Proc. Nat. Acad. Sci. U.S.* **53**, 1161.

Nishimura, S., Jones, D. S. & Khorana, H. G. 1965 *J. Mol. Biol.* **13**, 302.

Nishimura, S., Jones, D. S., Ohtsuka, E., Hayatsu, H., Jacob, T. M. & Khorana, H. G. 1965 *J. Mol. Biol.* **13**, 283.

Rifkin, D., Hirsh, D., Rifkin, M. R. & Konisberg, W. 1966 *Cold Spr. Harb. Symp. Quant. Biol.* **31** 'The genetic code'.

Salas, M., Smith, M. A., Stanley, W. M., Wahba, A. J. & Ochoa, S. 1965 *J. Biol. Chem.* **240**, 3988.

Sarabhai, A. S., Stretton, A. O. W., Brenner, S. & Bolle, A. 1964 *Nature, Lond.* **201**, 13.

Smith, J. D., Abelson, J. N., Clark, B. F. C., Goodman, H. M. & Brenner, S. 1966 *Cold Spr. Harb. Symp. Quant. Biol.* **31**, 'The genetic code'.

Smith, M. A., Salas, M., Stanley, W. M., Wahba, A. J. & Ochoa, S. 1966 *Proc. Nat. Acad. Sci. U.S.* **55**, 141.

Söll, D., Ohtsuka, E., Jones, D. S., Lohrmann, R., Hayatsu, H., Nishimura, A. & Khorana, H. G. 1965 *Proc. Nat. Acad. Sci. U.S.* **54**, 1378.

Söll, D., Jones, D. S., Ohtsuka, E., Faulkner, R. D., Khorana, H. G., Cherayil, J. D., Hampel, A. & Bock, R. M. 1966 *J. Mol. Biol.* **19**, 556.

Stanley, W. M. Jr., Salas, M., Wahba, A. J. & Ochoa, S. 1966 *Proc. Nat. Acad. Sci. U.S.* **56**, 290.

Streisinger, G., Okada, Y., Terzaghi, E., Emrich, J., Tsugita, A. & Inouye, M. 1966 *Cold Spr. Harb. Symp. Quant. Biol.* **31**, 'The genetic code'.

Sundararajan, T. A. & Thach, R. E. 1966 *J. Mol. Biol.* **19**, 74.

Takanami, M. & Yan, Y. 1965 *Proc. Nat. Acad. Sci. U.S.* **54**, 1450.

Terzaghi, E., Okada, Y., Streisinger, G., Emrich, J., Inouye, M. & Tsugita, A. 1966 *Proc. Nat. Acad. Sci. U.S.* **56**, 500.

Thach, R. E., Dewey, K. F., Brown, J. C. & Doty, P. 1966 *Science*, **153**, 416.

Trupin, J., Rottman, F., Brimacombe, R., Leder, P., Bernfield, M. & Nirenberg, M. 1965 *Proc. Nat. Acad. Sci. U.S.* **53**, 807.

Tsugita, A. 1962 *J. Mol. Biol.* **5**, 284.

von Ehrenstein, G. 1966 *Cold Spr. Harb. Symp. Quant. Biol.* **31**, 'The genetic code'.

Waller, J. P. 1963 *J. Mol. Biol.* **7**, 483.

Watson, J. D. 1965 *The molecular biology of the gene.* New York: Benjamin Inc.

Webster, R. E., Engelhardt, D. L. & Zinder, N. D. 1966 *Proc. Nat. Acad. Sci. U.S.* **55**, 155.

Weigert, M. & Garen, A. 1965 *Nature, Lond.* **206**, 992.

Wittmann, H. G. & Wittmann-Liebold, B. 1966 *Cold Spr. Harb. Symp. Quant. Biol.* **31**. 'The genetic code'.

Yanofsky, C., Carlton, B. C., Guest, J. R., Helinski, D. R. & Henning, U. 1964 *Proc. Nat. Acad. Sci. U.S.* **51**, 266.

Yanofsky, C. 1966 *Cold Spr. Harb. Symp. Quant. Biol.* **31**. 'The genetic code'.