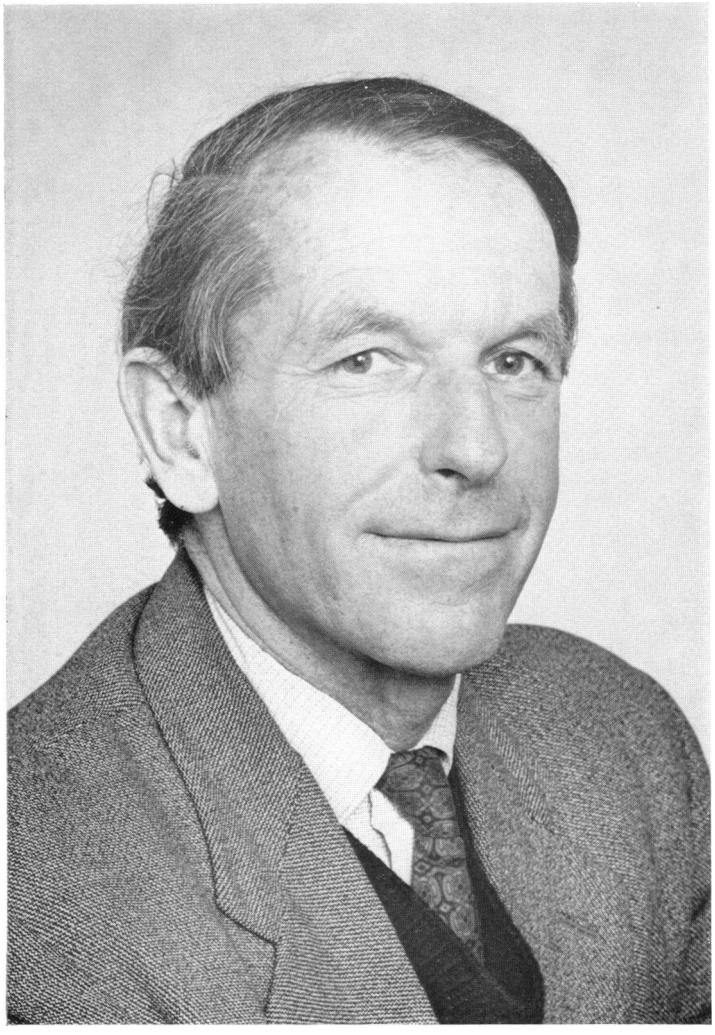
The Biochemical Journal, Vol. 124, *No.* 5



Dr F. SANGER, C.B.E., F.R.S.

Biochem. J. (1971) **124**, 833–843
Printed in Great Britain

Nucleotide Sequences in Bacteriophage Ribonucleic Acid

THE EIGHTH HOPKINS MEMORIAL LECTURE

BY F. SANGER*

Delivered at a Meeting of the Biochemical Society on 23 April 1971 at University College London, Gower Street, London WC1H 0AH, U.K.

I feel very honoured to have been invited to give this Lecture in memory of Frederick Gowland Hopkins. Although I only knew Hopkins for a short time towards the end of his life, I worked for many years in the Biochemistry Department at Cambridge, which he built up, and I feel I owe much to the spirit and outlook of that Department. These were largely due to Hopkins's own influence and personality, and in particular to his enthusiasm for research into the unknown fields of biochemistry—research that was carried out under conditions of friendly co-operation between scientists of differing personality and outlook.

In his Linacre Lecture Hopkins (1938) wrote: 'With all the events that can be directly observed in living cells, including those truly remarkable occurrences involved in cell growth and division, cell differentiation in development, and the like, which it has been the privilege of classical cytology and especially experimental cytology to reveal, there must be associated molecular events as varied and in some way as controlled as are the visible events themselves.'

Although today few scientists would question Hopkins's assertion of the importance of an underlying chemical mechanism at the basis of all biological events in living cells, it was in 1938 a somewhat unorthodox and prophetic view, and illustrates not only his enthusiasm for biochemistry but also his ability to foresee the directions in which it was leading.

One of the most important developments in biochemistry in the last 30 years, and one that has brought us much closer to Hopkins's ideal, has been the realization of the central role played by the nucleic acids in biology, and in particular their capacity—as genes and messenger RNA—to carry biological information in the form of specific sequences of nucleotide residues. A study of these sequences is therefore of particular interest, and in this Lecture I shall give an account of our studies on the sequences in the RNA of the bacteriophage R17.

* Address: Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

Bacteriophage R17

The RNA bacteriophages are some of the simplest organisms that exist. Bacteriophage R17 is composed of an RNA molecule 3300 residues long contained in a protein capsule. This capsule contains two proteins. The main component is the 'coat protein'; the complete sequence of its 129 amino acids has been determined by Weber (1967) and is shown in Fig. 3. There are approximately 180 molecules of this protein per bacteriophage particle and there is one molecule of a second protein, known as the 'A protein', whose function is not exactly known though it is probably involved in the maturation of the virus (Steitz, 1968). When the RNA enters the host cell (*Escherichia coli*) it catalyses the production of three proteins—the coat protein, the A protein and the 'replicase' (or 'synthetase'), which is responsible for copying the RNA. Since no DNA is involved in the replication, the RNA acts both as a gene that carries all the information for the organism and as a messenger RNA on which the proteins are synthesized. Fig. 1 shows the order in which the three protein cistrons are arranged on the RNA. This order was finally established during the course of the work described below (Jeppesen, Steitz, Gesteland & Spahr, 1970). The nucleotide sequences in the RNA are translated into amino acid sequences through the genetic code (Table 3). This code had been established largely by binding experiments and, although there seemed little doubt that it was correct, a direct confirmation by chemical means was desirable, and one of the purposes of the present study was to establish a direct relationship between the sequence of a protein and the part of the messenger RNA that was coding for it. It was also of interest to determine which codons were used in a particular situation, but it seemed likely that the most interesting sequences would be those that did not actually code for amino acids but would be expected to carry information for the starting and stopping of protein chains and for the detailed control of protein synthesis, since such control is most probably

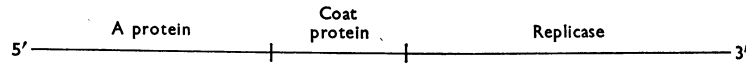


Fig. 1. Order of the protein cistrons in bacteriophage R17 RNA.

responsible for the 'truly remarkable occurrences' that Hopkins referred to.

Fractionation methods

The RNA of the bacteriophage was a single chain of 3300 residues, and a study of its sequence was a formidable problem that could only be attacked by the development of new techniques. Progress in sequence analysis, both in the protein and nucleic acid fields, has usually depended very much on the progress in fractionation techniques. In our early work we were therefore particularly concerned with trying to develop new techniques for the fractionation of oligonucleotides, and we were particularly interested in the use of small-scale paper methods of fractionation (chromatography and ionophoresis).

The first RNA whose complete sequence was determined was the alanine transfer RNA, which was studied by Holley *et al.* (1965). In this work the partial enzymic digests were fractionated on columns of DEAE-cellulose. This was a very efficient method for fractionating oligonucleotides and it was possible to deduce a complete sequence. These methods have now been applied to a number of other transfer RNA species, but it was considered that in order to study larger RNA molecules it was necessary to use simpler and more rapid methods. Fractionation of oligonucleotides by ionophoresis on paper, which had been used extensively for peptide fractionation, was not very successful, since the larger nucleotides tended to streak badly and were not well resolved. However, it was found possible to use certain modified papers and thin-layer systems and to develop two-dimensional techniques for the fractionation of relatively small oligonucleotides (Sanger, Brownlee & Barrell, 1965; Brownlee & Sanger, 1969).

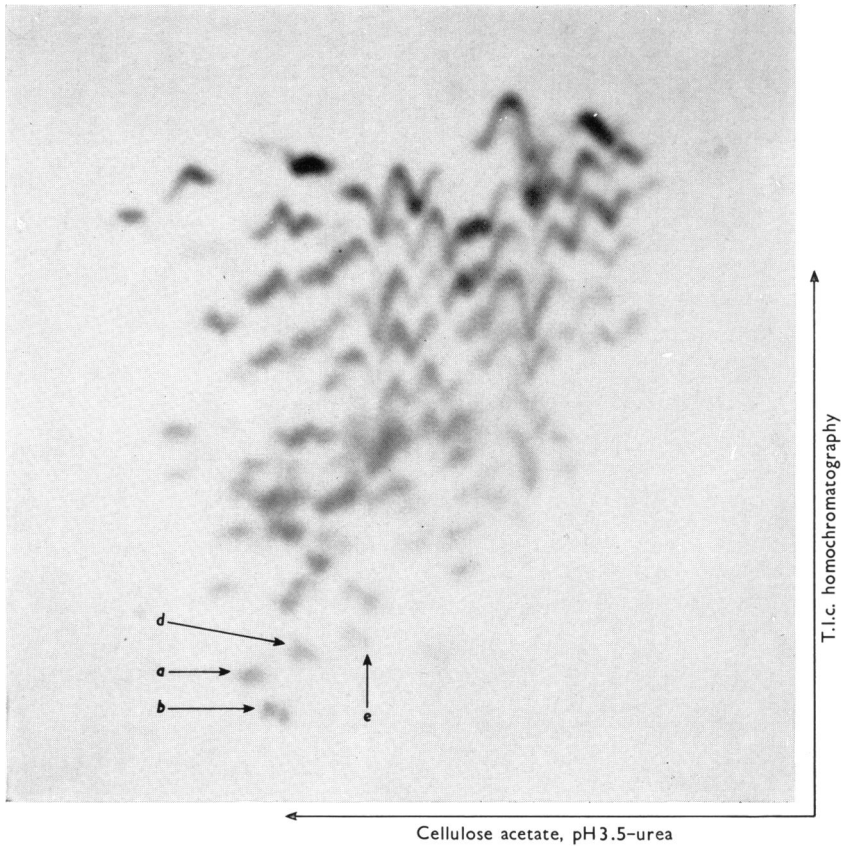
In using micro techniques involving this type of fractionation it is necessary to have a very sensitive method for detection and determination of the oligonucleotides. Previously this has been done by making use of the absorption at 260nm; however, it was insufficiently sensitive for use with complicated mixtures fractionated on two-dimensional systems. For this reason we have used RNA labelled with ^{32}P. This is a convenient isotope that can be obtained in high specific radioactivity, and every mononucleotide residue contains one phosphorus atom. The ^{32}P-labelled RNA was usually prepared from bacterial sources by growing organisms in a medium containing [^{32}P]phosphate.

Micro methods were also developed for the analysis and sequence determination of the labelled oligonucleotides, and with these techniques it was possible to deduce the complete sequence of ribosomal 5S RNA, which contains 120 residues (Brownlee & Sanger, 1967; Brownlee, Sanger & Barrell, 1968).

Ribonuclease T_1 digest of bacteriophage R17 RNA

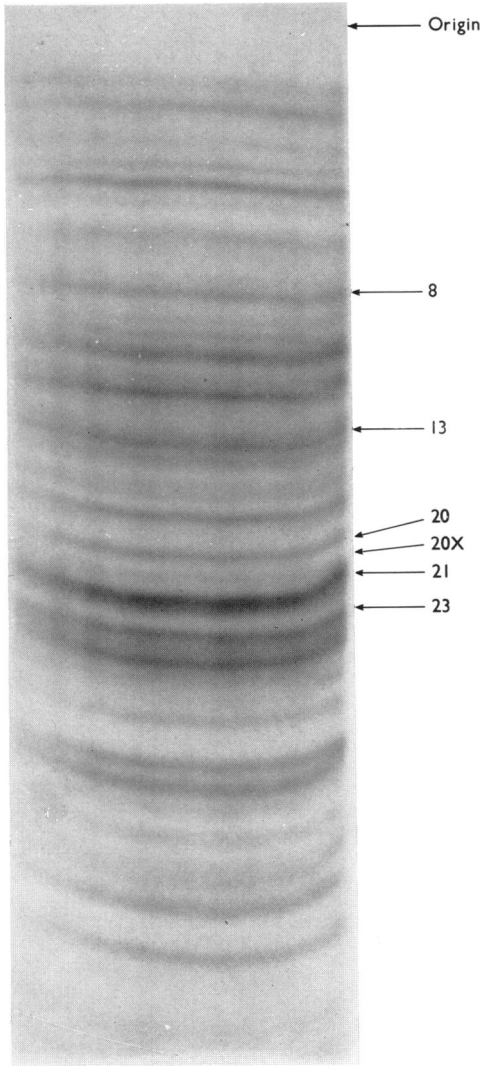
It seemed unlikely at the time that the above methods could be applied to a molecule as large as the bacteriophage R17 RNA. However, useful preliminary results could be obtained by applying a two-dimensional fractionation technique to a ribonuclease T_1 digest of the bacteriophage R17 RNA, as shown in Plate 1 (Jeppesen, 1971; Adams, Jeppesen, Sanger & Barrell, 1969). Ribonuclease T_1 is the most useful enzyme for carrying out digests of RNA because it is specific, splitting only at the 5′-bond following Gp residues. It thus gives rise to a mixture of nucleotides that are terminated at their 3′-end by -Gp. In the 'fingerprint' shown in Plate 1 the first dimension was carried out by using high-voltage ionophoresis at pH 3.5 on cellulose acetate. Although on ionophoresis on paper large oligonucleotides usually streak badly, on cellulose acetate nucleotides of almost any size move as sharp well-defined spots; we have therefore almost always used this as the first dimension in the systems we have developed. For the second dimension in Plate 1 we used a special type of ion-exchange chromatography on thin layers of DEAE-cellulose. In this method, which is known as 'homochromatography', a mixture of non-radioactive nucleotides was used to develop the chromatogram. The different oligonucleotides formed a series of fronts on which the corresponding radioactive nucleotides were carried down the chromatogram and separated (Brownlee & Sanger, 1969).

Plate 1 represents a recent experiment by Jeppesen (1971) in which the resolution is considerably better than in some of the earlier experiments (Adams *et al.* 1969). In a ribonuclease T_1 digest all the Gp residues are split, so that a large number of oligonucleotides is produced and it would not be expected that all of these could be resolved on the simple two-dimensional system. Resolution



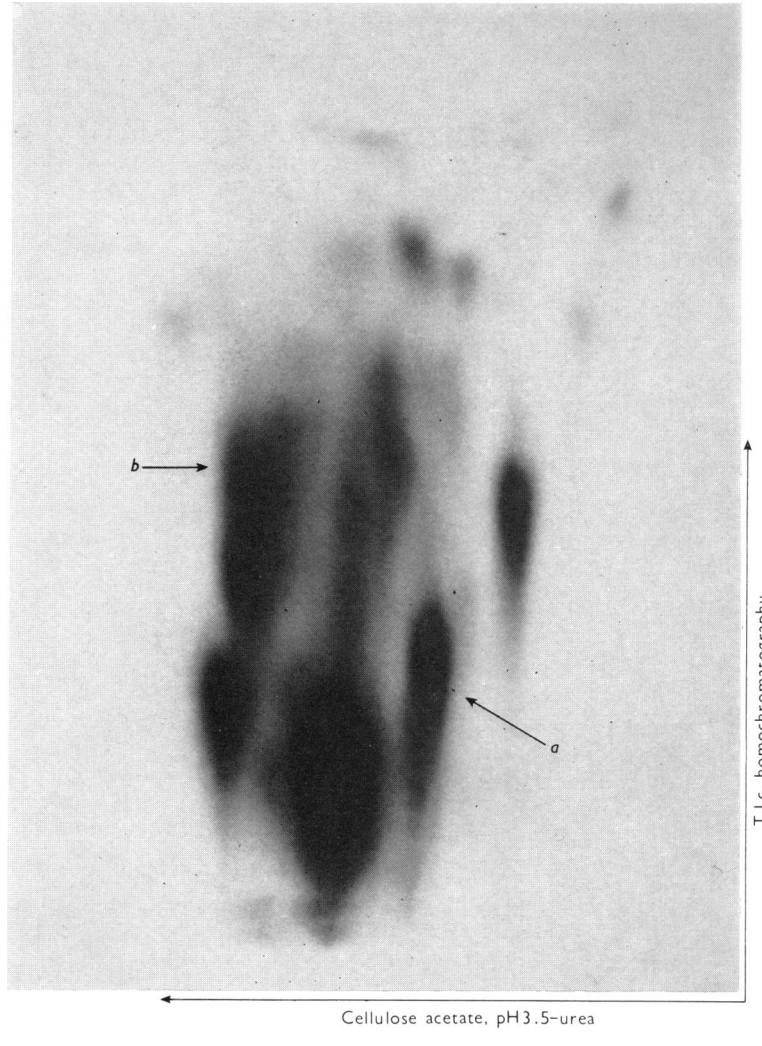
EXPLANATION OF PLATE 1

Radioautograph of a two-dimensional fractionation of a ribonuclease T_1 digest of ^{32}P-labelled bacteriophage R17 RNA (Jeppesen, 1971).



EXPLANATION OF PLATE 2

Radioautograph of a polyacrylamide-gel electrophoresis of a partial ribonuclease T_1 digest of ^{32}P-labelled bacteriophage R17 RNA.



EXPLANATION OF PLATE 3

Purification of band 20X from the polyacrylamide-gel fractionation (Plate 2) of a partial ribonuclease T_1 digest of ^{32}P-labelled bacteriophage R17 RNA.

on the homochromatography dimension is largely dependent on the size of the nucleotides; the small ones move more rapidly and are not well resolved. However, there are only a few of the larger nucleotides (i.e. more than about 15 residues long) that move slowly on the system, and these are separated as discrete spots that can be eluted from the thin-layer plate and subjected to sequence analysis. The sequence of many of these has now been determined (Jeppesen, 1971).

Initially we were interested to see whether any of these fragments could be derived from the coat-protein cistron. This cistron represents only about 10% of the total RNA and, as an initial screening experiment, the fragments were subjected to digestion with ribonuclease A, which splits at Cp and Up bonds and gives rise to products of the type $(Ap)_nNp$, where N is C, U or G. Such products could be rapidly separated on a one-dimensional ionophoresis system and gave a useful initial characterization of each nucleotide. Table 1 shows the products of pancreatic ribonuclease digestion of some of the larger fragments isolated from the 'fingerprint' shown in Plate 1.

Fig. 2 illustrates the way in which one can test whether a particular oligonucleotide could fit into the coat-protein cistron. From the known amino acid sequence of the coat protein it is possible to write down the possible nucleotide sequence of the cistron, by making use of the genetic code. Fig. 2 shows a part of the amino acid sequence, residues 83–97, together with the corresponding nucleotide sequence. Owing to the degeneracy of the code there are many ambiguities in the sequence. From the possible RNA sequence one can then list the possible pancreatic-ribonuclease digestion products. Having made a complete list of pancreatic products for the whole protein, as in Fig. 2, it is then possible to check through for any oligonucleotide whose pancreatic-ribonuclease digestion products are known. Thus spot *e* (Plate 1) contains A-A-A-Up (Table 1). In the part of the cistron shown in Fig. 2 the only possible A-A-A-Up sequence is in position 86–87, but spot *e* also contains A-A-Cp and there is no possible ribonuclease T_1 product that could contain both A-A-A-Up and A-A-Cp.

Table 1. *Ribonuclease A digestion products of some large oligonucleotides from a ribonuclease T_1 digest of bacteriophage R17 RNA*

Spot (Plate 1)	Ribonuclease A digestion products
a	A-A-Up_2, A-A-Cp, A-Up, Gp, Cp_3, Up_6
b	A-A-A-A-Up, A-A-Cp, A-Up_3, Gp, Cp_4, Up_5
d	A-A-A-Up, A-A-Up_2, A-Cp, Gp, Cp_2, Up_3
e	A-A-A-Up, A-A-Cp, A-Up, Gp, Cp_5, Up_3

Similarly, by screening through the whole coat-protein sequence in this way it could be shown that the combination of pancreatic-ribonuclease digestion products of spot *e* could not occur in any part of the coat-protein cistron. It was also shown that spots *b* and *d* were not derived from the coat-protein cistron. However, the products obtained from spot *a* could be accommodated in a possible ribonuclease T_1 product corresponding to positions 90–96. Spot *a* was therefore worth investigating further and its complete sequence was determined (Table 2). It was found to correspond exactly, through the genetic code, to the above amino acid sequence. As any heptapeptide sequence would be expected to be unique in a molecule of this size, it seemed fairly certain from the exact correspondence of the two sequences that the ribonuclease T_1 oligonucleotide was in fact derived from this part of the coat-protein cistron.

Various methods have been developed for determining the detailed sequence of oligonucleotides when only small amounts of radioactive material are available. These methods usually depend on further degradation with other enzymes. The two methods that have been used most extensively in this work are as follows.

(1) The carbodi-imide method. The water-soluble carbodi-imide reagent [*N*-cyclohexyl-*N'*-(β-morpholinyl-4-ethyl)carbodi-imide-methyl toluene-*p*-sulphonate] of Gilham (1962) reacts specifically with guanosine and uridine residues and the product formed with uridine is not split by ribonuclease A. Thus if a ribonuclease T_1 oligonucleotide is treated with the reagent and then digested with ribonuclease A, splitting occurs only at the Cp residues.

(2) Ribonuclease U_2 (Arima, Uchida & Egami, 1968). This enzyme reacts specifically with purine residues so that in a ribonuclease T_1 oligonucleotide only the Ap residues are split.

Table 2 shows the products obtained from nucleotide *a* by the above methods and illustrates how the complete sequence could be worked out.

Partial digests of bacteriophage R17 RNA

The study of the large oligonucleotides from a ribonuclease T_1 digest is somewhat limited, and in order to go further it was necessary to try and isolate longer fragments from the bacteriophage R17 RNA. The most effective method for fractionating relatively large oligonucleotides is electrophoresis on polyacrylamide gel, and experiments by Gould (1967) had shown that discrete fragments could be separated from partial digests of ribosomal RNA by this method. By using a similar approach on flat slabs of polyacrylamide gel, it was possible to obtain a fractionation of large fragments

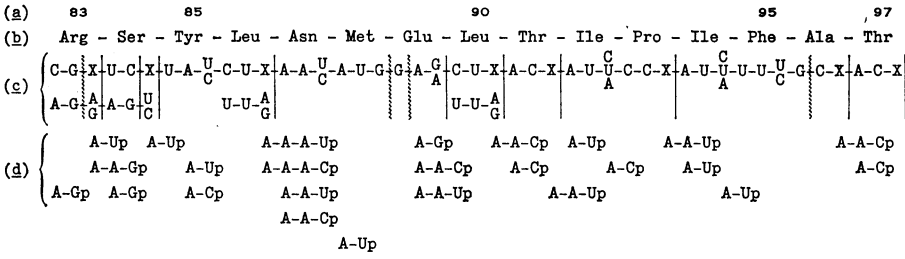


Fig. 2. Method for screening large ribonuclease T_1 digestion products of bacteriophage R17 RNA to test whether they can be derived from the coat-protein cistron. (*a*) Position of amino acid residues in coat protein; (*b*) part of the amino acid sequence of coat protein; (*c*) corresponding nucleotide sequence predicted from the genetic code; (*d*) possible ribonuclease A digestion products of bacteriophage R17 RNA. | indicates possible sites of cleavage with ribonuclease T_1; ‡ indicates obligatory sites of cleavage with ribonuclease T_1.

Table 2. *Determination of the sequence of nucleotide* a *from a ribonuclease T_1 digest of bacteriophage R17 RNA*

Ribonuclease A products	Ribonuclease A products after blocking with carbodi-imide reagent (CD products)	Ribonuclease U_2 products (U_2 products)
A-A-Up(2)	(A-A-Ů,Ů)A-A-Cp	Ap
A-A-Cp	(A-Ů,$Ů_2$)Cp	A-Ap
A-Up	(A-A-Ů,$Ů_3$)Cp	U-U-A-Ap
Gp	Gp	U-U-Ap
Cp	Cp	C-U-Ap
Up		(U_2,C_2)A-Ap
		(U_2,C_2)Ap
		(U_4,C)Gp

CD products ... ...
U_2 products ... ...
Sequence of oligonucleotide *a*: G-A-A-U-U-A-A-C-U-A-U-U-C-C-A-A-U-U-U-U-C-Gp
Amino acid sequence in positions 89–95 of coat protein: Glu - Leu - Thr - Ile - Pro - Ile - Phe

from a very limited ribonuclease T_1 partial digest of bacteriophage R17 RNA (Adams *et al.* 1969). Plate 2 shows a radioautograph of a polyacrylamide-gel slab on which a limited partial digest of bacteriophage R17 RNA was separated. The presence of well-defined bands was rather unexpected. The average size is probably about 50–100 residues, suggesting that something like 5–10% of the Gp residues had been split. If this splitting had occurred equally at each Gp residue then a very complex mixture should have resulted. The fact that discrete bands could be seen indicated that the digestion must have been relatively specific, certain bonds being split almost completely whereas others were resistant. The products present were now in the size range for which methods of sequence analysis had already been devised.

In initial experiments the RNA present in each band was subjected to complete digestion with ribonuclease T_1 and the products were fractionated on the homochromatography two-dimensional system. The larger products from this were then screened to see whether they could be derived from the coat-protein cistron. Band 21 was found to give nucleotide *a*, which had already been shown to be present in the coat-protein cistron, and it was therefore subjected to partial degradation and the complete sequence determined. This is shown in Fig. 3 as the nucleotide sequence corresponding to residues 81–100 of the protein sequence. It can be seen that this nucleotide sequence is related, by the genetic code, exactly to the corresponding amino acid sequence. This correspondence left no doubt that the fragment was a part of the coat-protein cistron and offered a direct chemical confirmation for the correctness of the genetic code—at least as far as the amino acids in this part of the molecule were concerned.

Four other fragments from the polyacrylamide-gel fractionation have now been identified as com-

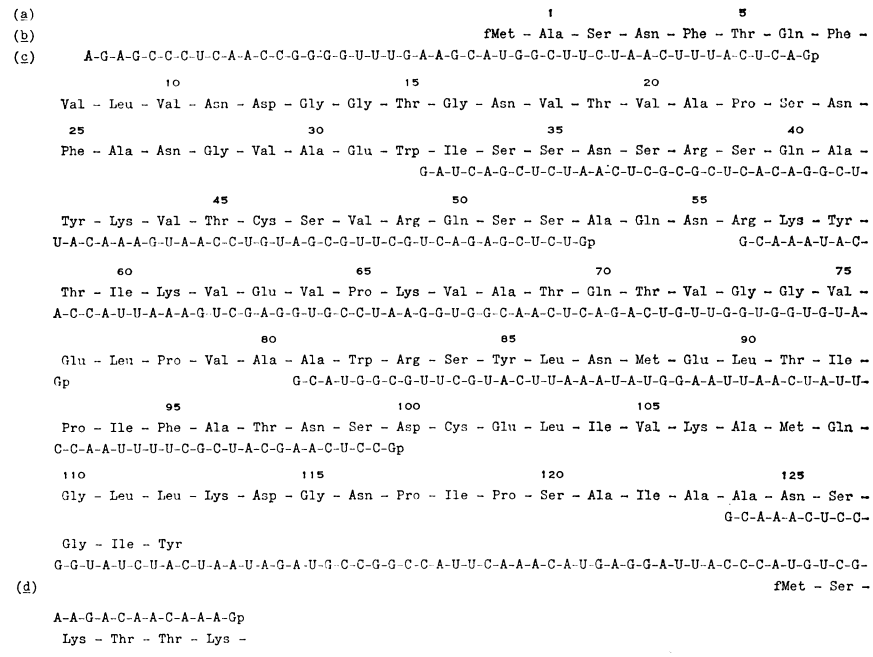


Fig. 3. Amino acid sequence of the coat protein and nucleotide sequence found in the corresponding cistron of bacteriophage R17 RNA. (*a*) Position of amino acid residues in coat protein; (*b*) amino acid sequence of coat protein (Weber, 1967); (*c*) nucleotide sequence in bacteriophage R17 RNA; (*d*) amino acid sequence of replicase.

ponents of the coat-protein cistron, and their sequence is included in Fig. 3. All of them are of a similar size (50–60 residues) and therefore run close together on the polyacrylamide-gel ionophoresis.

Nucleotide loops

Band 20X component was present in rather small amounts and was heavily contaminated with other fragments, including the nucleotide from band 21. In order to purify it it was subjected to refractionation on the two-dimensional system by using homochromatography (Plate 3). Mobility on polyacrylamide gel and on the homochromatography system depends largely on size, so that it was rather surprising to find a separation of spots on the homochromatography. The main component was largely the oligonucleotide from band 21 and was 57 residues long, but two smaller components (*a* and *b*) were present. The complete nucleotide sequence of these two is given in Fig. 4. Spot *a* component was 31 residues long and its sequence was related to the amino acids in positions 55–66 of the coat-protein. The sequence of spot *b* component was related to the amino acid sequence immediately following (67–76), with a single Gp residue missing. Thus it appears that, whereas these two fragments migrate together on the polyacrylamide gel in a position that would indicate a size of 60 residues, in the homochromatography system they have separated as two smaller fragments. (The fractionation on the two-dimensional system is carried out in 8 M-urea solution, whereas no urea is present during the polyacrylamide-gel ionophoresis.) The explanation for this anomaly can be understood if the structure of the two fragments is written in the form of a 'hairpin loop', as in Fig. 5(*a*). Such a loop forms a large number of base pairs and would be expected to be a relatively stable structure. It thus seems probable that a structure of this type is present in the intact RNA and that splitting with ribonuclease T_1 occurs at the positions arrowed, a single Gp residue being split out from the end of the loop. The two fragments are still held together during the polyacrylamide-gel ionophoresis, but on refractionation in 8 M-urea are separated. Similar loops also appear to be present in the oligonucleotides from bands 20 and 21 (Fig. 5).

It seems probable that these loops are present in the intact RNA and are probably concerned with the folding of the RNA in the virus particle. It is possible that they may be concerned also in other functions of the messenger RNA, such as the control

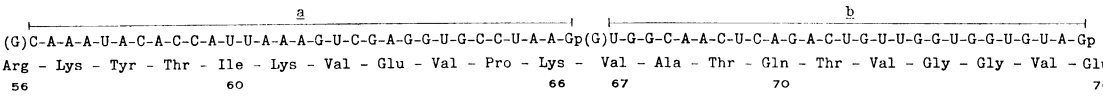


Fig. 4. Sequences of oligonucleotides in spots *a* and *b* from the purification (Plate 3) of band 20X from the polyacrylamide-gel fractionation (Plate 2) of a partial ribonuclease T_1 digest of ^{32}P-labelled bacteriophage R17 RNA, showing their relationship to the amino acid sequence in positions 56–76 of the coat protein.

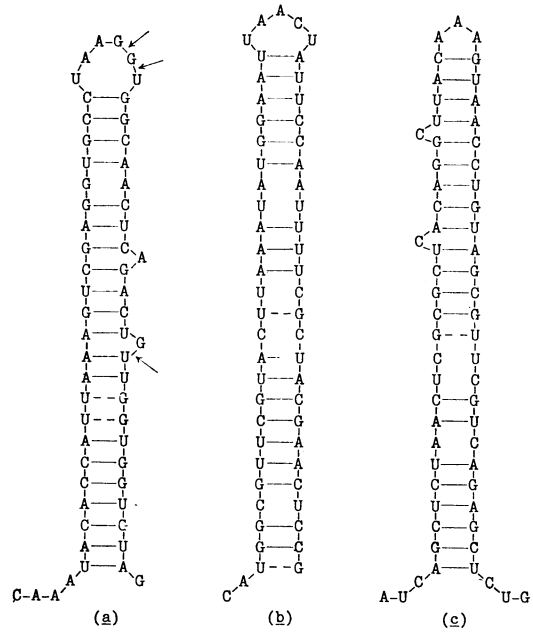


Fig. 5. Fragments of bacteriophage R17 RNA written in the form of 'hairpin loops'. (*a*) Band 20X component; (*b*) band 21 component; (*c*) band 20 component. Arrows indicate sites of cleavage with ribonuclease T_1 under conditions for partial digestion.

of protein synthesis, but at present this is speculative. However, the results do indicate that a messenger RNA—at least in a virus—may have a considerable secondary structure. The sequence of a messenger RNA may thus be determined, not only by the sequence of amino acids in the protein for which it codes, but also by its own secondary structure. In view of the ambiguities in the code these two functions need not necessarily be conflicting, as it is possible to alter the sequence—particularly in the third positions of codons—to obtain maximum base-pairing without affecting the coding specificity. Since protein biosynthesis depends on the recognition of codons by base-pairing with transfer RNA, the messenger RNA must be single-stranded during translation. It therefore appears that the ribosomes must be able to unfold the loops during translation. One could thus predict that there is a constant change in secondary structure of the RNA as the ribosome moves along the messenger RNA. Such changes may be of importance in the control of protein synthesis and initiation. For instance, in this way different initiation sites might be exposed for different periods.

Ribosomal binding sites

An alternative approach for studying specific sequences in the bacteriophage R17 RNA has been used by Steitz (1969), who prepared a complex of ^{32}P-labelled bacteriophage R17 RNA with ribosomes under conditions where translation of the RNA did not take place. The ribosomes were bound on to the sites where translation starts, which correspond to the *N*-termini of the three proteins. The complex was then subjected to digestion with pancreatic ribonuclease. The binding sites were, however, protected by the presence of the ribosome, and a sequence of about 30 residues was left intact. These fragments were isolated and their complete nucleotide sequences determined (Fig. 6). Sequence 1 contains a nucleotide sequence at its 3′-end that corresponds to the amino acid sequence at the *N*-terminus of the coat protein; it was therefore identified as the binding site for initiation of coat-protein synthesis. It also contains an A-U-G sequence, which codes for the initiating formylmethionine transfer RNA. The other two sequences also each contained an A-U-G sequence in the initiating site and they were identified from a knowledge of the *N*-terminal residues of the A protein and the replicase respectively.

Although there are many A-U-G sequences present in the bacteriophage R17 RNA, only the above three are recognized by the ribosomes as initiation sites. It was therefore expected that the three initiation sites would have further sequences in common, which would represent the recognition signal. However, apart from the A-U-G sequence, the three sequences did not have anything obvious in common, and so it is not clear what features the ribosome does recognize in these particular A-U-G sequences. One possibility is that different sites may be recognized by different ribosomes or different factors. In this respect it may be significant

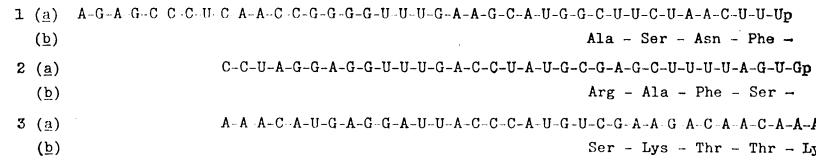


Fig. 6. Ribosomal binding sites of bacteriophage R17 RNA (Steitz, 1969). 1, Coat protein; 2, A protein; 3, replicase. (*a*) Nucleotide sequences of binding sites; (*b*) *N*-terminal amino acid sequences in proteins. That for the coat protein was determined by Weber (1967); those for the A protein and replicase were largely deduced from the nucleotide sequence, except for the *N*-terminal residues.

that the ribosomal binding sites for the coat protein and the A protein have the common sequence G-G-U-U-U-G-A. On the other hand the binding site for the replicase does not contain this sequence. It may therefore be that the coat and the A protein are recognized by one factor and the replicase by another.

Another possibility is that the translation system may recognize a three-dimensional structure rather than a specific nucleotide sequence. As present concepts of protein structure and function have shown the importance of secondary structure in protein reactions, it may be that this also plays an important part in initiation—especially as proteins are also likely to be involved. Recent experiments by Lodish (1970, 1971) have indicated that changes in secondary structure do lead to changes in the pattern of initiation. The sequence for the initiation of coat-protein synthesis can be written in the form of a base-paired loop in which the A-U-G sequence is at the end of the loop (Fig. 7), and could thus constitute a very specific structure that could be recognized. However, the other two sequences cannot be written in a similar form.

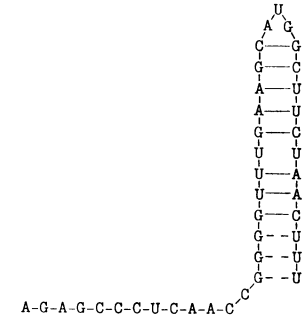


Fig. 7. Nucleotide sequence of the coat-protein-binding site of bacteriophage R17 RNA (Fig. 6) written in the form of a loop.

The termination signal

One of the fragments from the polyacrylamide-gel fractionation of the partial digest (band 23 in Plate 2) was found to have a nucleotide sequence that corresponded to the *C*-terminal amino acid sequence of the coat protein (Nichols, 1970). This was followed by the sequence U-A-A-U-A-G. The triplets U-A-A and U-A-G are two of the 'nonsense' codons that have been found not to code for an amino acid (Table 3), and if present in a messenger RNA (as in 'ochre' and 'amber' mutations respectively) lead to termination of the peptide chain. The presence of these two terminators in sequence was unexpected, but may represent a mechanism to ensure complete termination. Clearly if the termination were not complete a larger unphysiological protein would be produced, which might cause disastrous biological effects. The complete sequence of the nucleotide in band 23 is shown in Fig. 8, which also shows the sequence for the initiation site of the replicase protein. These two sequences have the sequence A-A-A-C-A-U-G in common and therefore most probably overlap one another. In this way it is possible to write the complete RNA sequence between the two protein cistrons. It may be noted that there is a third termination sequence, U-G-A, which is in phase with the other two and may act as a third line of defence against non-termination. There are about 30 residues between the two cistrons. The exact function of these is uncertain, but it seems probable that they are involved in the control of protein synthesis, and probably in determining how much of the replicase is produced at certain stages in the infectious cycle. The detailed unravelling of the message contained in these residues is not at present possible, but it is hoped that a study of more sequences of this type will reveal what principles are involved. Following the two termination signals, U-A-A and U-A-G, there is an A-U-G triplet. If this were to act as an initiation signal it would lead to the production of the pentapeptide Pro-Ala-Ile-Gln-Thr. There is no experimental evidence for the production of this peptide or that the A-U-G triplet acts as an initiation signal, so that it seems most probable that its presence in this position is merely fortuitous.

However, we cannot rule out the possibility that small amounts of this peptide might be produced and have some interesting biological function.

The coat-protein cistron

Figs. 3 and 9 summarize the present position with regard to the nucleotide sequence in the coat-protein cistron. In Fig. 9 the three looped structures are shown as solid lines, their relative positions being deduced from the amino acid sequence. There are about 18 residues between them in each case. The broken lines show positions where the nucleotide sequence is unknown. There are two fairly long stretches, near the two termini, that have been tentatively put in as loops. To what extent it will be possible to complete the sequence of the cistron is uncertain: most of the bands from the polyacrylamide-gel fractionation have been screened but their detailed sequence has not been worked out. It appears that to be able to continue very much further with the study of the sequence in the cistron it will be necessary to develop improved techniques, and particularly techniques for the fractionation of larger degradation products. The work on fragments from the coat protein was greatly facilitated by a knowledge of the amino acid sequence, and it will clearly be more difficult to deduce sequences in other areas of the molecule.

Although previous studies on the genetic code have shown which triplets could code for each amino acid, they did not decide which particular codon was used in practice. In Table 3 the number of times each codon has been found during the above studies is shown. It can be seen that the code is degenerate—i.e. many different codons are used for the same amino acid. For instance, all four possible codons for each of valine and threonine have been found and all the six possible ones for serine. On the other hand, on each of the four times that tyrosine has been detected it has been coded for by the U-A-C triplet rather than by the U-A-U triplet. Similarly glutamine has been coded for four times by the C-A-G triplet. At present these findings are probably not statistically significant enough to warrant any generalization, but they may indicate that certain codons are used more frequently than others.

The 5′-terminus

At the 5′-terminus of the RNA there is a triphosphate residue that can be detected by the production of pppGp on alkaline hydrolysis. This was also present in band 13 from the polyacrylamide-gel fractionation. The complete sequence of the band 13 component has been determined by Adams & Cory (1970) and is shown in Fig. 10 written in the form of two base-paired loops. The cistron nearest to the 5′-end on the RNA is the A-protein cistron, and since there is no overlap between the sequence of the band 13 component and the A-protein-binding site it was concluded that the 74 residues in Fig. 10 were not translated. Band 8 from the polyacrylamide-gel fractionation also contains the 5′-end. Preliminary results by U. Rensing & B. G.

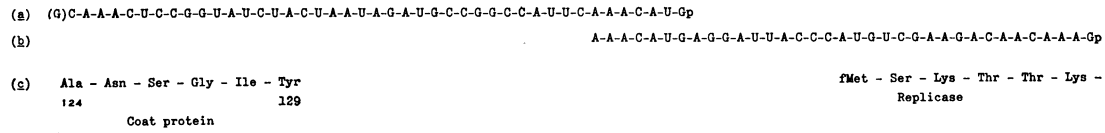


Fig. 8. Sequence of the region between the cistrons for the coat protein and replicase in bacteriophage R17 RNA (Nichols, 1970). (*a*) Nucleotide sequence of the band 23 component from the polyacrylamide-gel electrophoresis (Plate 2) of a partial ribonuclease T_1 digest of the RNA; (*b*) nucleotide sequence of the replicase-binding site (Fig. 6); (*c*) corresponding amino acid sequences.

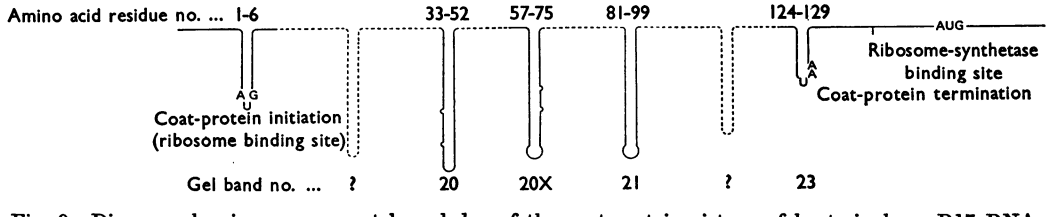


Fig. 9. Diagram showing our present knowledge of the coat-protein cistron of bacteriophage R17 RNA. Continuous lines represent known nucleotide sequences.

Table 3. *The genetic code*

The numbers in parentheses after the amino acid residues indicate the number of times that particular codon has been found in bacteriophage R17 RNA.

Second letter ... First letter	Genetic code U	C	A	G	Third letter
U	Phe (2)	Ser (4)	Tyr	Cys (1)	U
U	Phe (1)	Ser (2)	Tyr (4)	Cys	C
U	Leu (2)	Ser (1)	—	—	A
U	Leu	Ser (3)	—	Trp (1)	G
C	Leu	Pro (1)	His	Arg (2)	U
C	Leu	Pro	His	Arg (2)	C
C	Leu	Pro (1)	Gln	Arg (1)	A
C	Leu	Pro	Gln (4)	Arg	G
A	Ile (3)	Thr (4)	Asn (1)	Ser (1)	U
A	Ile (2)	Thr (2)	Asn (4)	Ser (3)	C
A	Ile	Thr (2)	Lys (3)	Arg	A
A	Met (1)	Thr (1)	Lys (3)	Arg	G
G	Val (2)	Ala (4)	Asp	Gly (3)	U
G	Val (1)	Ala	Asp	Gly	C
G	Val (2)	Ala (3)	Glu (1)	Gly	A
G	Val (2)	Ala	Glu (1)	Gly	G

Barrell (unpublished work) show that it overlaps the A-protein-binding site and indicate that there are about 130 residues before the A-U-G triplet that initiates the synthesis of the A protein. Similar results have been obtained by De Wachter, Vandenberghe, Merregaert, Contreras & Fiers (1971) with the closely related bacteriophage MS2. Using a novel pulse-labelling technique, Billeter, Dahlberg, Goodman, Hindley & Weissmann (1969) have determined a sequence of 175 residues at the 5′-end of bacteriophage Qβ RNA. No possible initiation site was present in at least the first 62 residues.

The function of these untranslated residues is not certain, but it is again probable that they are involved in the control of protein synthesis. During replication the replicase copies the bacteriophage RNA (the + strand), starting from its 3′-end, to give the complementary structure (the − strand). It then copies the − strand, starting at its 3′-end. Since the sequence at the 3′-end of the − strand is determined by the sequence at the 5′-end of the + strand, both ends of the + strand must have a signal for the replicase to recognize. It would be expected that the signal would be the same for both strands, and on this basis Adams & Cory (1970) have suggested that the short sequence -C-C-A-C-C-C-OH at the 3′-end, which is complementary to the sequence pppG-G-G-U-G-G- at the 5′-end, is the recognition site. An analogous mechanism could not, however, account for the recognition site in bacteriophage Qβ RNA (Goodman, Billeter, Hindley & Weissmann, 1970).

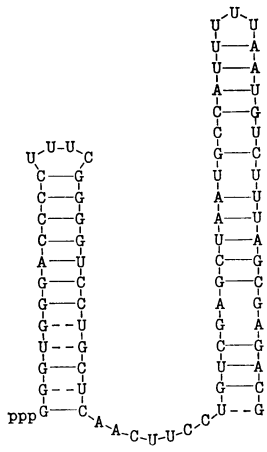


Fig. 10. Nucleotide sequence at the 5′-end of bacteriophage R17 RNA (Adams & Cory, 1970).

Sequences in related bacteriophages

Bacteriophage R17 belongs to a closely related group, some other members of which have recently been studied. Nichols & Robertson (1971) have studied the bacteriophage f2, and Fiers and colleagues (Min Jou, Haegeman & Fiers, 1971; De Wachter *et al.* 1971) have studied bacteriophage MS2. The coat proteins of bacteriophages MS2 and R17 are identical, whereas that of bacteriophage f2 differs by only one amino acid. Nichols & Robertson (1971) have determined the nucleotide

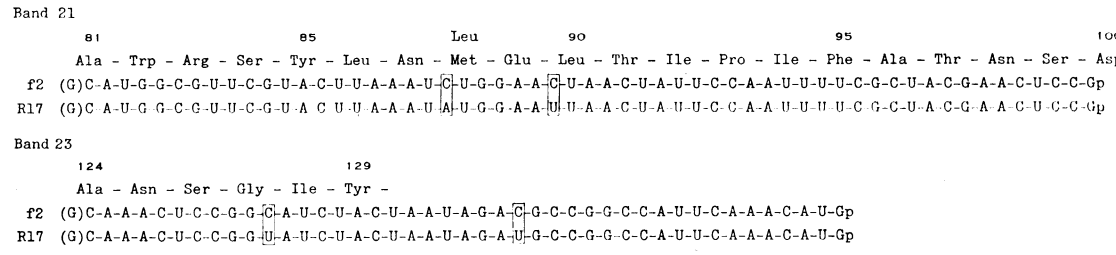


Fig. 11. Comparison of nucleotide sequences in the coat-protein cistrons of bacteriophages f2 and R17 (Nichols & Robertson, 1971). Nucleotides enclosed within solid lines indicate positions at which sequences differ. Residue 88 of the coat protein is leucine in bacteriophage f2 and methionine in bacteriophage R17. Otherwise the amino acid sequences of the two proteins are identical.

sequences of fragments from bacteriophage f2 RNA corresponding to bands 21 and 23 of bacteriophage R17 RNA, and these are shown in Fig. 11. As an approximation, it appears that about one in 20 nucleotides differs between the two phages. This is in contrast with a difference of one amino acid in 129 in the coat protein, and most of the nucleotide changes are in positions that do not lead to amino acid changes. Of interest is the fact that the triplet following the termination codons, which is A-U-G in the case of bacteriophage R17 RNA, is changed to A-C-G in the case of bacteriophage f2 RNA. Therefore in the case of bacteriophage f2 no pentapeptide—as discussed above—could be produced as A-C-G could not act as an initiation codon. This raises the possibility that a small change of this type might account for certain differences in the biological behaviour of the two bacteriophages. However, such a suggestion is very speculative at present.

Recently Ling (1971) has determined the sequence at the 5′-terminus of bacteriophage f2 RNA and De Wachter *et al.* (1971) have determined the corresponding sequence in bacteriophage MS2 RNA. In all three cases the sequences of the first 74 residues are identical. This is in contrast with the differences found in the sequences in the coat-protein cistron, and suggests that the function of the 5′-terminus is a very specific one that depends uniquely on nucleotide sequences.

Conclusion

In this Lecture I have illustrated how these new micro methods of sequence analysis can be used to study a messenger RNA, and thus the way in which information is transferred from the DNA to the completed organism. This is clearly one of the most important problems in biology. In this work we have been able to confirm the genetic code by a direct chemical comparison of a nucleic acid sequence and a protein sequence, and have been able to see how the code is used by a particular organism. Some information has been obtained about the signals that are responsible for the starting and stopping of protein chains, and we have also determined other sequences in regions of the RNA that do not code for protein chains. Although at present we do not fully understand what these sequences mean, they will probably prove to be the most interesting and to carry information for the control of protein synthesis and determine how much of a particular protein is produced at a given time. This is a major problem when one considers differentiation of a higher organism, and it is hoped that if further methods for studying messenger RNA species, and possibly DNA species, in higher organisms can be developed it may be possible to begin to understand how a specific sequence of bases can be converted into a living organism, or, as Hopkins (1936) put it: 'Chemical differentiation underlies, or is associated with, morphological and functional differentiation, and to learn exactly what is the nature of such association is a fascinating task ahead.'

REFERENCES

Adams, J. M. & Cory, S. (1970). *Nature, Lond.*, **227**, 570.
Adams, J. M., Jeppesen, P. G. N., Sanger, F. & Barrell, B. G. (1969). *Nature, Lond.*, **223**, 1009.
Arima, T., Uchida, T. & Egami, F. (1968). *Biochem. J.* **106**, 609.
Billeter, M. A., Dahlberg, J. E., Goodman, H. M., Hindley, J. & Weissmann, C. (1969). *Nature, Lond.*, **224**, 1083.
Brownlee, G. G. & Sanger, F. (1967). *J. molec. Biol.* **23**, 337.
Brownlee, G. G. & Sanger, F. (1969). *Eur. J. Biochem.* **11**, 395.
Brownlee, G. G., Sanger, F. & Barrell, B. G. (1968). *J. molec. Biol.* **34**, 379.
De Wachter, R., Vandenberghe, A., Merregaert, J., Contreras, R. & Fiers, W. (1971). *Proc. natn. Acad. Sci. U.S.A.* **68**, 585.
Gilham, P. T. (1962). *J. Am. chem. Soc.* **84**, 687.

Goodman, H. M., Billeter, M. A., Hindley, J. & Weissmann, C. (1970). *Proc. natn. Acad. Sci. U.S.A.* **67**, 921.
Gould, H. J. (1967). *J. molec. Biol.* **29**, 307.
Gould, H. J., Pinder, J. C., Matthews, H. R. & Gordon, A. H. (1969). *Analyt. Biochem.* **29**, 1.
Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R. & Zamir, A. (1965). *Science, N.Y.*, **147**, 1462.
Hopkins, F. G. (1936). *Science, N.Y.*, **84**, 258.
Hopkins, F. G. (1938). *Lancet*, i, 1201.
Jeppesen, P. G. N. (1971). *Biochem. J.* **124**, 357.
Jeppesen, P. G. N., Steitz, J. A., Gesteland, R. F. & Spahr, P. F. (1970). *Nature, Lond.*, **226**, 230.
Ling, V. (1971). *Biochem. biophys. Res. Commun.* **42**, 82.
Lodish, H. F. (1970). *J. molec. Biol.* **50**, 689.
Lodish, H. F. (1971). *J. molec. Biol.* **56**, 627.
Min Jou, W., Haegeman, G. & Fiers, W. (1971). *FEBS Lett.* **13**, 105.
Nichols, J. L. (1970). *Nature, Lond.*, **225**, 147.
Nichols, J. L. & Robertson, H. D. (1971). *Biochim. biophys. Acta*, **228**, 676.
Sanger, F., Brownlee, G. G. & Barrell, B. G. (1965). *J. molec. Biol.* **13**, 373.
Steitz, J. A. (1968). *J. molec. Biol.* **33**, 923.
Steitz, J. A. (1969). *Nature, Lond.*, **224**, 957.
Weber, K. (1967). *Biochemistry, Easton*, **6**, 3144.