

## GENERAL NATURE OF THE GENETIC CODE FOR PROTEINS

By DR. F. H. CRICK, (F.R.S.) LESLIE BARNETT, DR. S. BRENNER  
and DR. R. J. WATTS-TOBINMedical Research Council Unit for Molecular Biology,  
Cavendish Laboratory, Cambridge

THERE is now a mass of indirect evidence which suggests that the amino-acid sequence along the polypeptide chain of a protein is determined by the sequence of the bases along some particular part of the nucleic acid of the genetic material. Since there are twenty common amino-acids found throughout Nature, but only four common bases, it has often been surmised that the sequence of the four bases is in some way a code for the sequence of the amino-acids. In this article we report genetic experiments which, together with the work of others, suggest that the genetic code is of the following general type:

(a) A group of three bases (or, less likely, a multiple of three bases) codes one amino-acid.

(b) The code is not of the overlapping type (see Fig. 1).

(c) The sequence of the bases is read from a fixed starting point. This determines how the long sequences of bases are to be correctly read off as triplets. There are no special 'commas' to show how to select the right triplets. If the starting point is displaced by one base, then the reading into triplets is displaced, and thus becomes incorrect.

(d) The code is probably 'degenerate'; that is, in general, one particular amino-acid can be coded by one of several triplets of bases.

## The Reading of the Code

The evidence that the genetic code is not overlapping (see Fig. 1) does not come from our work, but from that of Wittmann<sup>1</sup> and of Tsugita and Fraenkel-Conrat<sup>2</sup> on the mutants of tobacco mosaic virus produced by nitrous acid. In an overlapping triplet code, an alteration to one base will in general change three adjacent amino-acids in the polypeptide chain. Their work on the alterations produced in the protein of the virus show that usually only one amino-acid at a time is changed as a result of treating the ribonucleic acid (RNA) of the virus with nitrous acid. In the rarer cases where two amino-acids are altered (owing presumably to two separate deaminations by the nitrous acid on one piece of RNA), the altered amino-acids are not in adjacent positions in the polypeptide chain.

Brenner<sup>3</sup> had previously shown that, if the code were universal (that is, the same throughout Nature), then all overlapping triplet codes were impossible. Moreover, all the abnormal human haemoglobins studied in detail<sup>4</sup> show only single amino-acid changes. The newer experimental results essentially rule out all simple codes of the overlapping type.

If the code is not overlapping, then there must be some arrangement to show how to select the correct triplets (or quadruplets, or whatever it may be) along the continuous sequence of bases. One obvious suggestion is that, say, every fourth base is a 'comma'. Another idea is that certain triplets make 'sense', whereas others make 'nonsense', as in the comma-free

codes of Crick, Griffith and Orgel<sup>5</sup>. Alternatively, the correct choice may be made by starting at a fixed point and working along the sequence of bases three (or four, or whatever) at a time. It is this possibility which we now favour.

## Experimental Results

Our genetic experiments have been carried out on the *B* cistron of the  $r_{II}$  region of the bacteriophage *T4*, which attacks strains of *Escherichia coli*. This is the system so brilliantly exploited by Benzer<sup>6,7</sup>. The  $r_{II}$  region consists of two adjacent genes, or 'cistrons', called cistron *A* and cistron *B*. The wild-type phage will grow on both *E. coli B* (here called *B*) and on *E. coli K12* ( $\lambda$ ) (here called *K*), but a phage which has lost the function of either gene will not grow on *K*. Such a phage produces an *r* plaque on *B*. Many point mutations of the genes are known which behave in this way. Deletions of part of the region are also found. Other mutations, known as 'leaky', show partial function; that is, they will grow on *K* but their plaque-type on *B* is not truly wild. We report here our work on the mutant *P 13* (now re-named *FC 0*) in the *B1* segment of the *B* cistron. This mutant was originally produced by the action of proflavin<sup>8</sup>.

We<sup>9</sup> have previously argued that acridines such as proflavin act as mutagens because they add or delete a base or bases. The most striking evidence in favour of this is that mutants produced by acridines are seldom 'leaky'; they are almost always completely lacking in the function of the gene. Since our note was published, experimental data from two sources have been added to our previous evidence: (1) we have examined a set of 126  $r_{II}$  mutants made with acridine yellow; of these only 6 are leaky (typically about half the mutants made with base analogues are leaky); (2) Streisinger<sup>10</sup> has found that whereas mutants of the lysozyme of phage *T4* produced by base-analogues are usually leaky, all lysozyme mutants produced by proflavin are negative, that is, the function is completely lacking.

If an acridine mutant is produced by, say, adding a base, it should revert to 'wild-type' by deleting a base. Our work on revertants of *FC 0* shows that it usually

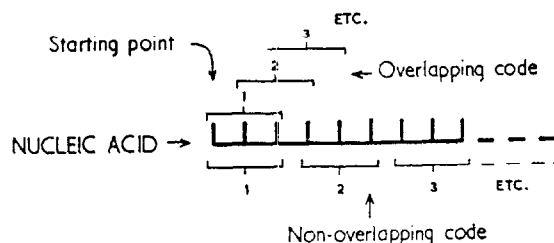


Fig. 1. To show the difference between an overlapping code and a non-overlapping code. The short vertical lines represent the bases of the nucleic acid. The case illustrated is for a triplet code

reverts not by reversing the original mutation but by producing a second mutation at a nearby point on the genetic map. That is, by a 'suppressor' in the same gene. In one case (or possibly two cases) it may have reverted back to true wild, but in at least 18 other cases the 'wild type' produced was really a double mutant with a 'wild' phenotype. Other workers<sup>11</sup> have found a similar phenomenon with  $r_{II}$  mutants, and Jinks<sup>12</sup> has made a detailed analysis of suppressors in the  $h_{III}$  gene.

The genetic map of these 18 suppressors of  $FC 0$  is shown in Fig. 2, line *a*. It will be seen that they all fall in the  $B1$  segment of the gene, though not all of them are very close to  $FC 0$ . They scatter over a region about, say, one-tenth the size of the  $B$  cistron. Not all are at different sites. We have found eight sites in all, but most of them fall into or near two close clusters of sites.

In all cases the suppressor was a non-leaky  $r$ . That is, it gave an  $r$  plaque on  $B$  and would not grow on  $K$ . This is the phenotype shown by a complete deletion of the gene, and shows that the function is lacking. The only possible exception was one case where the suppressor appeared to back-mutate so fast that we could not study it.

Each suppressor, as we have said, fails to grow on  $K$ . Reversion of each can therefore be studied by the same procedure used for  $FC 0$ . In a few cases these mutants apparently revert to the original wild-type, but usually they revert by forming a double mutant. Fig. 2, lines *b-g*, shows the mutants pro-

duced as suppressors of these suppressors. Again all these new suppressors are non-leaky  $r$  mutants, and all map within the  $B1$  segment for one site in the  $B2$  segment.

Once again we have repeated the process on two of the new suppressors, with the same general results, as shown in Fig. 2, lines *i* and *j*.

All these mutants, except the original  $FC 0$ , occurred spontaneously. We have, however, produced one set (as suppressors of  $FC 7$ ) using acridine yellow as a mutagen. The spectrum of suppressors we get (see Fig. 2, line *h*) is crudely similar to the spontaneous spectrum, and all the mutants are non-leaky  $r$ 's. We have also tested a (small) selection of all our mutants and shown that their reversion-rates are increased by acridine yellow.

Thus in all we have about eighty independent  $r$  mutants, all suppressors of  $FC 0$ , or suppressors of suppressors, or suppressors of suppressors of suppressors. They all fall within a limited region of the gene and they are all non-leaky  $r$  mutants.

The double mutants (which contain a mutation plus its suppressor) which plate on  $K$  have a variety of plaque types on  $B$ . Some are indistinguishable from wild, some can be distinguished from wild with difficulty, while others are easily distinguishable and produce plaques rather like  $r$ .

We have checked in a few cases that the phenomenon is quite distinct from 'complementation', since the two mutants which separately are phenotypically  $r$ , and together are wild or pseudo-wild,

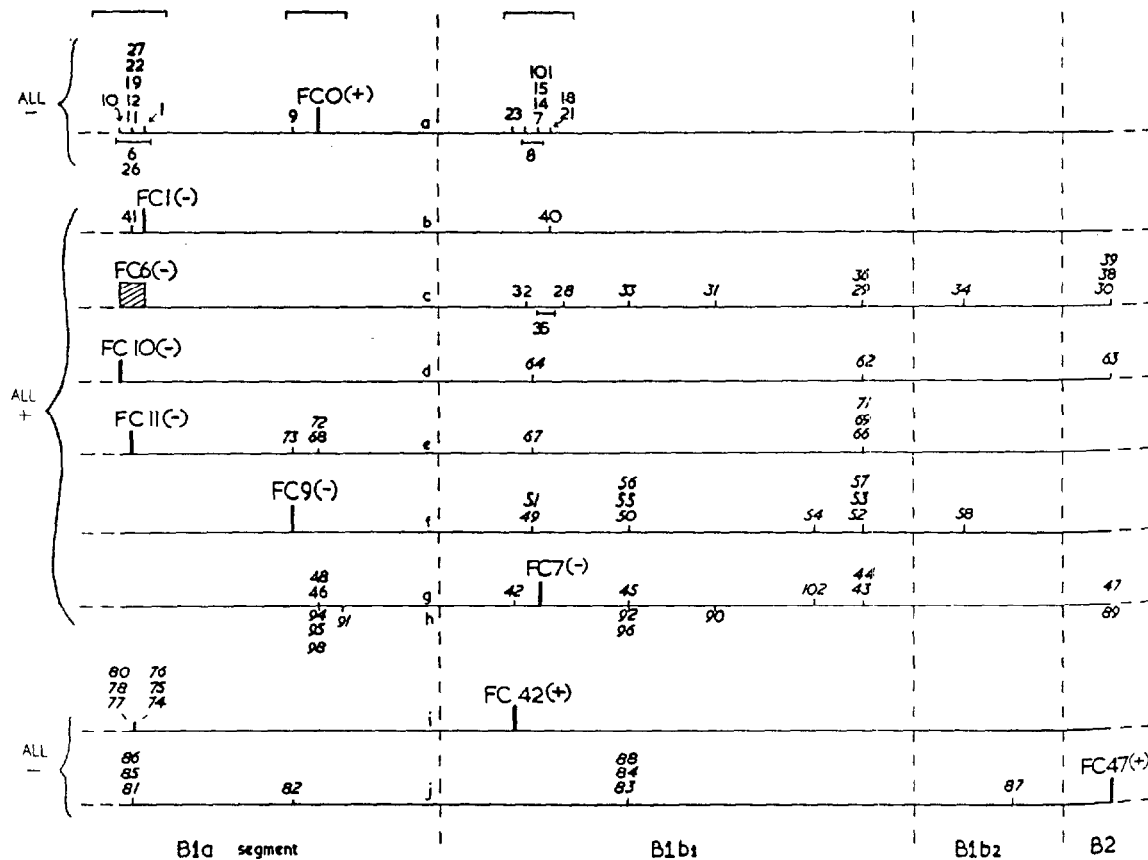


Fig. 2. A tentative map—only very roughly to scale—of the left-hand end of the  $B$  cistron, showing the position of the  $FC$  family of mutants. The order of sites within the regions covered by brackets (at the top of the figure) is not known. Mutants in italics have only been located approximately. Each line represents the suppressors picked up from one mutant, namely, that marked on the line in bold figures



the region to the right of the *FC 21* cluster. Thus we predict that a combination of a + with a - will be wild or pseudo-wild if it involves a → shift, but that such pairs involving a ← shift will be phenotypically *r* if the arrow crosses one or more of the forbidden places, since then an unacceptable triplet will be produced.

Table 2. DOUBLE MUTANTS OF THE TYPE (+ WITH -)

	<i>FC 41</i>	<i>FC 0</i>	<i>FC 40</i>	<i>FC 42</i>	<i>FC 58*</i>	<i>FC 63</i>	<i>FC 38</i>
<i>FC 1</i>	W	W	W	W	W	W	W
<i>FC 86</i>	W	W	W	W	W	W	W
<i>FC 9</i>	<i>r</i>	W	W	W	W	W	W
<i>FC 82</i>	<i>r</i>	W	W	W	W	W	W
<i>FC 21</i>	<i>r</i>	W	W	W	W	W	W
<i>FC 88</i>	<i>r</i>	<i>r</i>	W	W	W	W	W
<i>FC 87</i>	<i>r</i>	<i>r</i>	<i>r</i>	<i>r</i>	W	W	W

W, wild or pseudo-wild phenotype; W, wild or pseudo-wild combination used to isolate the suppressor; *r*, *r* phenotype.  
\* Double mutants formed with *FC 58* (or with *FC 34*) give sharp plaques on *K*.

We have tested this prediction in the 28 cases shown in Table 2. We expected 19 of these to be wild, or pseudo-wild, and 9 of them to have the *r* phenotype. In all cases our prediction was correct. We regard this as a striking confirmation of our theory. It may be of interest that the theory was constructed before these particular experimental results were obtained.

#### Rigorous Statement of the Theory

So far we have spoken as if the evidence supported a triplet code, but this was simply for illustration. Exactly the same results would be obtained if the code operated with groups of, say, 5 bases. Moreover, our symbols + and - must not be taken to mean literally the addition or subtraction of a single base.

It is easy to see that our symbolism is more exactly as follows:

+ represents +*m*, modulo *n*  
- represents -*m*, modulo *n*

where *n* (a positive integer) is the coding ratio (that is, the number of bases which code one amino-acid) and *m* is any integral number of bases, positive or negative.

It can also be seen that our choice of reading direction is arbitrary, and that the same results (to a first approximation) would be obtained in whichever direction the genetic material was read, that is, whether the starting point is on the right or the left of the gene, as conventionally drawn.

#### Triple Mutants and the Coding Ratio

The somewhat abstract description given above is necessary for generality, but fortunately we have convincing evidence that the coding ratio is in fact 3 or a multiple of 3.

This we have obtained by constructing triple mutants of the form (+ with + with +) or (- with - with -). One must be careful not to make shifts

Table 3. TRIPLE MUTANTS HAVING A WILD OR PSEUDO-WILD PHENOTYPE

<i>FC (0 + 40 + 38)</i>
<i>FC (0 + 40 + 58)</i>
<i>FC (0 + 40 + 57)</i>
<i>FC (0 + 40 + 54)</i>
<i>FC (0 + 40 + 55)</i>
<i>FC (1 + 21 + 23)</i>

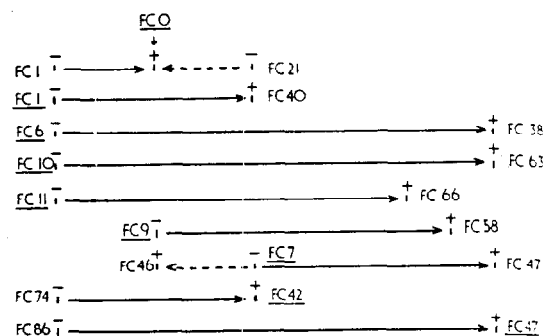


Fig. 4. A simplified version of the genetic map of Fig. 2. Each line corresponds to the suppressor from one mutant, here underlined. The arrows show the range over which suppressors have so far been found, the extreme mutants being named on the map. Arrows to the right are shown solid, arrows to the left dotted.

across the 'unacceptable' regions for the ← shifts, but these we can avoid by a proper choice of mutants.

We have so far examined the six cases listed in Table 3 and in all cases the triples are wild or pseudo-wild.

The rather striking nature of this result can be seen by considering one of them, for example, the triple (*FC 0* with *FC 40* with *FC 38*). These three mutants are, by themselves, all of like type (+). We can say this not merely from the way in which they were obtained, but because each of them, when combined with our mutant *FC 9* (-), gives the wild, or pseudo-wild phenotype. However, either singly or together in pairs they have an *r* phenotype, and will not grow on *K*. That is, the function of the gene is absent. Nevertheless, the combination of all three in the same gene partly restores the function and produces a pseudo-wild phage which grows on *K*.

This is exactly what one would expect, in favourable cases, if the coding ratio were 3 or a multiple of 3.

Our ability to find the coding ratio thus depends on the fact that, in at least one of our composite mutants which are 'wild', at least one amino-acid must have been added to or deleted from the polypeptide chain without disturbing the function of the gene-product too greatly.

This is a very fortunate situation. The fact that we can make these changes and can study so large a region probably comes about because this part of the protein is not essential for its function. That this is so has already been suggested by Champ and Benzer<sup>13</sup> in their work on complementation in the  $r_{II}$  region. By a special test (combined infection on *K*, followed by plating on *B*) it is possible to examine the function of the *A* cistron and the *B* cistron separately. A particular deletion, 1589 (see Fig. 5) covers the right-hand end of the *A* cistron and part of the left-hand end of the *B* cistron. Although 1589 abolishes the *A* function, they showed that it allows the *B* function to be expressed to a considerable extent. The region of the *B* cistron deleted by 1589 is that into which all our *FC* mutants fall.

#### Joining two Genes Together

We have used this deletion to re-inforce our idea that the sequence is read in groups from a fixed starting point. Normally, an alteration confined to the *A* cistron (be it a deletion, an acridine mutant, or any other mutant) does not prevent the expression of the *B* cistron. Conversely, no alteration within the *B* cistron prevents the function of the *A* cistron. This implies that there may be a region between the

two cistrons which separates them and allows their functions to be expressed individually.

We argued that the deletion 1589 will have lost this separating region and that therefore the two (partly damaged) cistrons should have been joined together. Experiments show this to be the case, for now an alteration to the left-hand end of the *A* cistron, if combined with deletion 1589, can prevent the *B* function from appearing. This is shown in Fig. 5. Either the mutant *P43* or *X142* (both of which revert strongly with acridines) will prevent the *B* function when the two cistrons are joined, although both of these mutants are in the *A* cistron. This is also true of *X142 S1*, a suppressor of *X142* (Fig. 5, case *b*). However, the double mutant (*X142* with *X142 S1*), of the type (+ with -), which by itself is pseudo-wild, still has the *B* function when combined with 1589 (Fig. 5, case *c*). We have also tested in this way the 10 deletions listed by Benzer<sup>7</sup>, which fall wholly to the left of 1589. Of these, three (386, 168 and 221) prevent the *B* function (Fig. 5, case *f*), whereas the other seven show it (Fig. 5, case *e*). We surmise that each of these seven has lost a number of bases which is a multiple of 3. There are theoretical reasons for expecting that deletions may not be random in length, but will more often have lost a number of bases equal to an integral multiple of the coding ratio.

It would not surprise us if it were eventually shown that deletion 1589 produces a protein which consists of part of the protein from the *A* cistron and part of that from the *B* cistron, joined together in the same polypeptide chain, and having to some extent the function of the undamaged *B* protein.

#### Is the Coding Ratio 3 or 6?

It remains to show that the coding ratio is probably 3, rather than a multiple of 3. Previous rough estimates<sup>10,14</sup> of the coding ratio (which are admittedly very unreliable) might suggest that the coding ratio is not far from 6. This would imply, on our theory, that the alteration in *FC* 0 was not to one base, but to two bases (or, more correctly, to an even number of bases).

We have some additional evidence which suggests that this is unlikely. First, in our set of 126 mutants produced by acridine yellow (referred to earlier) we have four independent mutants which fall at or

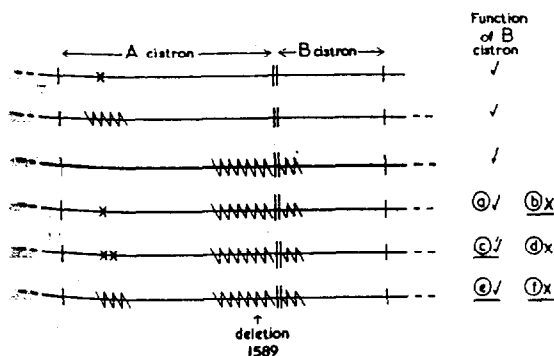


Fig. 5. Summary of the results with deletion 1589. The first two lines show that without 1589 a mutation or a deletion in the *A* cistron does not prevent the *B* cistron from functioning. Deletion 1589 (line 3) also allows the *B* cistron to function. The other cases, in some of which an alteration in the *A* cistron prevents the function of the *B* cistron (when 1589 is also present), are discussed in the text. They have been labelled (a), (b), etc., for convenience of reference, although cases (a) and (d) are not discussed in this paper. ✓ implies function; x implies no function

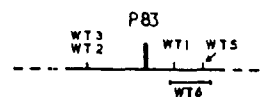


Fig. 6. Genetic map of *P83* and its suppressors, *WT1*, etc. The region falls within segment *B9a* near the right-hand end of the *B* cistron. It is not yet known which way round the map is in relation to the other figures

close to the *FC* 9 site. By a suitable choice of partners, we have been able to show that two are + and two are -. Secondly, we have two mutants (*X146* and *X225*), produced by hydrazine<sup>15</sup>, which fall on or near the site *FC* 30. These we have been able to show are both of type -.

Thus unless both acridines and hydrazine usually delete (or add) an even number of bases, this evidence supports a coding ratio of 3. However, as the action of these mutagens is not understood in detail, we cannot be certain that the coding ratio is not 6, although 3 seems more likely.

We have preliminary results which show that other acridine mutants often revert by means of close suppressors, but it is too sketchy to report here. A tentative map of some suppressors of *P83*, a mutant at the other end of the *B* cistron, in segment *B9a*, is shown in Fig. 6. They occur within a shorter region than the suppressors of *FC* 0, covering a distance of about one-twentieth of the *B* cistron. The double mutant *WT* (2 + 5) has the *r* phenotype, as expected.

#### Is the Code Degenerate?

If the code is a triplet code, there are 64 ( $4 \times 4 \times 4$ ) possible triplets. Our results suggest that it is unlikely that only 20 of these represent the 20 amino-acids and that the remaining 44 are nonsense. If this were the case, the region over which suppressors of the *FC* 0 family occur (perhaps a quarter of the *B* cistron) should be very much smaller than we observe, since a shift of frame should then, by chance, produce a nonsense reading at a much closer distance. This argument depends on the size of the protein which we have assumed the *B* cistron to produce. We do not know this, but the length of the cistron suggests that the protein may contain about 200 amino-acids. Thus the code is probably 'degenerate', that is, in general more than one triplet codes for each amino-acid. It is well known that if this were so, one could also account for the major dilemma of the coding problem, namely, that while the base composition of the DNA can be very different in different micro-organisms, the amino-acid composition of their proteins only changes by a moderate amount<sup>16</sup>. However, exactly how many triplets code amino-acids and how many have other functions we are unable to say.

#### Future Developments

Our theory leads to one very clear prediction. Suppose one could examine the amino-acid sequence of the 'pseudo-wild' protein produced by one of our double mutants of the (+ with -) type. Conventional theory suggests that since the gene is only altered in two places, only two amino-acids would be changed. Our theory, on the other hand, predicts that a string of amino-acids would be altered, covering the region of the polypeptide chain corresponding to the region on the gene between the two mutants. A good protein on which to test this hypothesis is

the lysozyme of the phage, at present being studied chemically by Dreyer<sup>17</sup> and genetically by Streisinger<sup>18</sup>.

At the recent Biochemical Congress at Moscow, the audience of Symposium I was startled by the announcement of Nirenberg that he and Matthaei<sup>18</sup> had produced polyphenylalanine (that is, a polypeptide all the residues of which are phenylalanine) by adding polyuridylic acid (that is, an RNA the bases of which are all uracil) to a cell-free system which can synthesize protein. This implies that a sequence of uracils codes for phenylalanine, and our work suggests that it is probably a triplet of uracils.

It is possible by various devices, either chemical or enzymatic, to synthesize polyribonucleotides with defined or partly defined sequences. If these, too, will produce specific polypeptides, the coding problem is wide open for experimental attack, and in fact many laboratories, including our own, are already working on the problem. If the coding ratio is indeed 3, as our results suggest, and if the code is the same throughout Nature, then the genetic code may well be solved within a year.

We thank Dr. Alice Orgel for certain mutants and for the use of data from her thesis, Dr. Leslie Orgel for many useful discussions, and Dr. Seymour Benzer for supplying us with certain deletions. We

are particularly grateful to Prof. C. F. A. Pantin for allowing us to use a room in the Zoological Museum, Cambridge, in which the bulk of this work was done.

- <sup>1</sup> Wittman, H. G., Symp. 1, Fifth Intern. Cong. Biochem., 1961, for refs. (in the press).
- <sup>2</sup> Tsugita, A., and Fraenkel-Conrat, H., *Proc. U.S. Nat. Acad. Sci.*, **46**, 636 (1960); *J. Mol. Biol.* (in the press).
- <sup>3</sup> Brenner, S., *Proc. U.S. Nat. Acad. Sci.*, **43**, 687 (1957).
- <sup>4</sup> For refs. see Watson, H. C., and Kendrew, J. C., *Nature*, **190**, 670 (1961).
- <sup>5</sup> Crick, F. H. C., Griffith, J. S., and Orgel, L. E., *Proc. U.S. Nat. Acad. Sci.*, **43**, 416 (1957).
- <sup>6</sup> Benzer, S., *Proc. U.S. Nat. Acad. Sci.*, **45**, 1607 (1959), for refs. to earlier papers.
- <sup>7</sup> Benzer, S., *Proc. U.S. Nat. Acad. Sci.*, **47**, 403 (1961); see his Fig. 3.
- <sup>8</sup> Brenner, S., Benzer, S., and Barnett, L., *Nature*, **182**, 983 (1958).
- <sup>9</sup> Brenner, S., Barnett, L., Crick, F. H. C., and Orgel, A., *J. Mol. Biol.*, **3**, 121 (1961).
- <sup>10</sup> Streisinger, G. (personal communication and in the press).
- <sup>11</sup> Feynman, R. P.; Benzer, S.; Freese, E. (all personal communications).
- <sup>12</sup> Jinks, J. L., *Heredity*, **16**, 153, 241 (1961).
- <sup>13</sup> Champe, S., and Benzer, S. (personal communication and in preparation).
- <sup>14</sup> Jacob, F., and Wollman, E. L., *Sexuality and the Genetics of Bacteria* (Academic Press, New York, 1961). Levinthal, C. (personal communication).
- <sup>15</sup> Orgel, A., and Brenner, S. (in preparation).
- <sup>16</sup> Sueoka, N., *Cold Spring Harb. Symp. Quant. Biol.* (in the press).
- <sup>17</sup> Dreyer, W. J., Symp. 1, Fifth Intern. Cong. Biochem., 1961 (in the press).
- <sup>18</sup> Nirenberg, M. W., and Matthaei, J. H., *Proc. U.S. Nat. Acad. Sci.*, **47**, 1588 (1961).