BACTERIOPHAGE T4 AND OUR PRESENT CONCEPT OF THE GENE

A. H. DOERMANN, UNIVERSITY OF WASHINGTON, SEATTLE, WASHINGTON

There is probably no geneticist who has been more critical than LEWIS J. STADLER in his interpretation of genetic experiments and more careful to make clear the distinction between experimental facts and the hypothetical evaluation of those facts. His classic paper on 'The Gene', which appeared in the November 19, 1954 issue of 'Science' as the published version of his Presidential Address to the American Society of Naturalists in 1953, is a most eloquent example of the rigorous and incisive analysis which he brought to bear on the conception of the gene at that time. Today, fifteen years later, STADLER's paper still stands out as a landmark in the history of genetics. It is no small honor to be given here the opportunity to pay tribute to that great geneticist.

The Committee for the Symposium asked me to speak on the role which studies of bacteriophages have played in the development of genetic concepts. I have taken this to mean neither a comprehensive review of phage genetics as it stands today, nor a detailed review of new experiments which might lead to changes in such concepts. It would be impossible here to enumerate every appropriate contribution, and I will restrict myself to reviewing what we have learned about 'gene structure' from experiments with bacteriophage T4. This is an old story, but it seems to be one which should be included in a symposium on genetic concepts, and more particularly a symposium honoring LEWIS J. STADLER. It should be understood, of course, that almost none of the work to be mentioned was totally independent of the work carried out with other organisms. My focus on T4 derives largely from the conviction that T4 has been a major contributor to our present concept of the gene.

THE CLASSICAL GENE

Perhaps the most appropriate starting point is STADLER's conception of the gene as expressed in his paper of 1954. It is a difficult concept to summarize, and I feel that this can probably best be accomplished by quoting several short passages from the paper. In defining the gene he writes "Operationally, the gene can be defined only as the smallest segment of the gene-string that can be shown to be consistently associated with the occurrence of a specific genetic effect. It cannot be defined as a single molecule, because we have no experimental operations that can be applied in actual cases to determine whether or not a given gene is a single molecule. It cannot be defined as an indivisible unit, because, although our definition provides that we will recognize as separate genes any determiners actually separated by crossing over or translocation, there is no experimental operation that can prove that further separation is impossible." In discussing the truism that our concept of the gene is entirely dependent upon the occurrence of gene mutations he points out "... that we have no positive criterion to identify mutations caused by a change within the gene, and that the alterations interpreted as gene mutations in experiments are merely the unclassified residue that cannot be proved to be due to other causes. The major objective in further investigations must be to develop such criterions." And again, pointing up the direction which future experiments must take, he wrote "The main purpose of this paper is to emphasize the unpleasant

fact that significant progress in our understanding of gene mutations requires the investigation of the mutation of specific genes."

THE rII STORY: DISSECTION OF THE CLASSICAL GENE

Our picture of the gene, then, as little as 15 years ago, included few details. A change in that situation awaited development of genetic systems in which a large number of distinguishable mutants could be used to analyse the fine structure of one, or perhaps several genes. In early 1954, almost the same time as the STADLER paper appeared, SEYMOUR BENZER (1955) was already taking advantage of an accidental discovery and evolving a system suitable for such an analysis. The discovery and development of this material is perhaps the most important breakthrough in T4 genetics, at least for our present concept of gene structure. Because that system is the foundation of much that will follow, it must be discussed at least briefly. HERSHEY and ROTMAN had, in 1948, described the so-called r-mutants of bacteriophage T2, mutants which make larger plaques with sharper margins than wild-type phages when plated against a member of the Escherichia coli B family of strains. They had also demonstrated that wild-type recombinants are usually obtained following mixed infection with two r strains of separate mutational origin. It was not, however, until six years later that an observation by BENZER made possible the development of the enormous potential of this genetic material. He noted that members of one set of these 4 mutants (now known as the rII set) are incapable of reproducing in E. coli strains which carry phage λ in the lysogenic state. For technical reasons BENZER turned from T2 to the closely related phage T4 and with that virus showed that it is feasible to score quantitatively among rII phages rII^+ particles with frequencies as low as 10^{-8}. Those r mutants unable to make plaques on λ-lysogens all fell into a single linked cluster which consisted of two categories defined by their complementation abilities. Any mutant falling into one group proved capable of giving normal yields of progeny when it infected a λ-lysogenic cell together with a mutant from the other group, while intragroup mixed infections gave no yield at all. Significantly all members of one group mapped to the right of a certain point while all members of the other group mapped to the left. BENZER recognized the importance of this material for analysing gene structure in detail, and by intensive application of the selective plating procedure and by supplementing his arsenal with the efficient technical developments of "spot testing" (BENZER 1955) and "deletion mapping" (BENZER 1961), he provided geneticists with a picture of two genes which has not been equalled as a mutational or a recombinational analysis of individual genes.

The whole BENZER approach was founded on his adopting operational definitions of the elements under study. Thus "the smallest element in the one-dimensional array that is interchangeable (but not divisible) by genetic recombination" was given the name 'recon'; "the smallest element that, when altered, can give rise to a mutant form of the organism" was referred to as the 'muton'; and the term 'cistron' was used to indicate "a map segment, corresponding to a function which is unitary as defined by the 'cis-trans' test applied to the heterocaryon" (BENZER 1957). Even the earliest experiments made clear that the cistron was divisible and could be analysed in terms of recons and mutons. Systemmatic use of these strict definitions made it possible for him to devise the strongest case for intragenic linearity (BENZER 1959); and from the large number of recons and mutons uncovered in both the rIIA and the rIIB cistrons he concluded that their numbers per cistron could not be far different from the number of nucleotide pairs in the DNA segment composing the functional unit (BENZER 1961).

Perhaps BENZER's most thorough effort was the analysis of the topography of these two cistrons, a study which included the mapping of

more than 2,400 revertible mutants. Among the 308 recons to which these mutants were assigned, some proved to be highly mutable. For example a single site in the B cistron accounted for 517 and one site in the A cistron for 292 of the approximately 1,600 spontaneous mutants mapped. On the other hand, many sites were encountered only once or twice, and if it was assumed that these represented a random sample of the classes which have the lowest order of mutability, he could estimate from a Poisson distribution how many more sites remained to be discovered. Using all the data from both spontaneous and induced mutations it was calculated that 120 undiscovered sites must exist in the two cistrons. The estimate, is, of course, a minimum estimate since sites of lower mutability than that of the single- and double-occurrence class cannot be excluded. Taking that difficulty into account, the analysis indicates that a minimum of 428 sites exist (308 discovered + 120 undiscovered). The number of nucleotide pairs in the rIIA + rIIB cistrons is probably less than ten times as high.

With this extraordinarily well-suited material BENZER was able to supply answers to some of the questions which LEWIS STADLER had posed in 1954, and in much greater detail than STADLER could have hoped. He was able to show, using operationally defensible definitions, that the gene in its functional sense cannot be defined as an indivisible unit, but rather that recombination goes on at many sites within its borders. Furthermore, he established that the intragenic structure is, in all likelihood, a linear one, a concept that is today no longer questioned. And his work even demonstrated that the sites identified within a gene are different from one another by proving extreme differences in their individual mutabilities. Yet all this might not have qualified the rII discovery as the most important breakthrough in T4 genetics. The further uses to which the system was subsequently put, especially for the study of gene structure, played an important part in establishing the role it had in the development of our present concept of the gene.

FRAME-SHIFT MUTATIONS AND THE GENERAL NATURE OF THE GENETIC CODE

One of the most far-reaching developments in which rII mutants played a crucial part is in the early development of the genetic code. In 1960 it was generally accepted that in some way the base sequence in DNA encoded the information which specified amino acid sequences. While the 64 combinations possible with three bases coding for one amino acid appeared excessive, a doublet code did not supply enough combinations for the number of natural amino acids known. Therefore a triplet code seemed necessary. At about the same time another feature of the code was already being clarified, namely that it was not an overlapping code. In the base sequence...ABCD..., if ABC codes for one amino acid, then the BCD in this sequence does not code for another. BRENNER (1957) by accepting the restriction of a universal code, had already shown that all overlapping triplet codes are impossible because, to account for all the amino acid sequences known already at that time would require appreciably more than 64 coding triplets. It was also shown by WITTMANN (1962) and by TSUGITA and FRAENKEL-CONRAT (1962) that chemical mutagens which presumably cause single-base changes in tobacco mosaic virus RNA bring about only single amino acid replacements and not replacement of two or three adjacent amino acids as would be expected from an overlapping code. To this information, CRICK, BARNETT, BRENNER, and WATTS-TOBIN (1961), making use of a small segment of the T4rIIB cistron, brought ingenious experimental evidence from which they concluded that the genetic code must be described as follows: (1) A group of three consecutive bases codes for one amino acid. (2) The code is not overlapping. (3) The sequence of bases is read from a fixed starting point. (4) The code is degenerate, that is, a particular amino acid may be coded by one of several triplets. It is the work which led to that picture of the genetic code which I would like now to summarize.

Earlier work had already brought BRENNER, BARNETT, CRICK and ORGEL (1961) to the conclusion that acridines act as mutagens because they promote insertion of an additional base in, or the deletion of a base from, the DNA molecule. In either case, the reading frame in a triplet code would be shifted at the site of change so that the distal triplets would now code for an entirely new set of amino acids. If both additions and deletions are possible they argued, an acridine-induced rII mutant resulting from an addition might be expected to revert to wild-type (or to a pseudo-wild phenotype) not necessarily by deleting that same base, but more frequently by deleting any base in the vicinity. In that case, rather than producing a protein with the entire terminal amino acid sequence changed, only the section coded by the addition-containing triplet, the deletion-containing triplet, and the segment between them should be altered. If the segment changed is not too long, then it might well prove to be a functional protein.

Experiments to test that idea were begun on the proflavine-induced rIIB mutant known as FC-0, and the first results already supported the hypothesis. Of 20 spontaneous revertants (able to grow on a λ-lysogen) at least 18 proved to be due to a second mutation which could be separated from the original FC-0, and which mapped in 8 recombinable sites close to the original mutation. Each of these suppressors of FC-0 was a non-leaky rII mutation with typical r plaque morphology on E. coli strain B and unable to grow on K(λ). Such rII mutants (first suppressors) could be studied by the same procedure used to investigate suppression of FC-0. Revertants of six suppressors were tested with similar results: the suppressors of suppressors were again non-leaky rII mutations located in the vicinity of the first suppressors. Revertants of two suppressors of suppressors were also tested with similar results. Assuming correctness of the original frame-shift idea, it was now possible to assign arbitrarily to the original mutation in FC-0 the sign (+), and then all first suppressors of it must be (-), and once again, suppressors of suppressors must then be (+). It should perhaps be emphasized that the (+) sign may designate either an addition or a deletion of a base, while the (-) sign simply indicates the opposite.

With this material available and in harmony with the hypothesis, further predictions could then be put to test. First, double mutants formed by recombining any (+) with another (+) or any (-) with another (-) should still be incapable of growth on K(λ) since no frame correction would be introduced into the genome. Four double (-) and 10 double (+) recombinants were made and all retained the r phenotype. It was also expected that any (+) should be able to suppress any (-) mutation. This simple expectation was not verified by the data: some (-) mutations failed to suppress certain (+) mutations. This unexpected result could, however, be explained by a further simple assumption. In those cases where a (+) and a (-) failed to suppress one another, the cause could be that the reading frame shift causes an unacceptable triplet to be formed. The location of such a point in the map was subsequently called a 'barrier'. If that explanation was correct it made a further clearcut prediction: any single-base frame shift made 'in the same direction' across the barrier site should also fail to cause suppression, while a shift in the opposite direction over that site would probably not form the same unacceptable triplet there. That prediction could be tested with the well distributed (+) and (-) mutations available. Thirty-four double mutants of (+) with (-) were constructed. Of these 9 retained the rII phenotype. The r and wild phenotypes are completely consistent with the occurrence of two (unidirectional) barriers, one to the right and one to the left of FC-0. These results gave further strong confirmation to the general notion that these r mutations represent addition or deletion frame shifts and that the mutual suppression of (+) and (-) mutations is caused by reestablishing a correct reading frame in the coding mechanism.

While all the foregoing arguments supported the idea of a general coding mechanism which reads groups of bases along a nucleic acid molecule, the number of bases which make up a code unit was not yet established. The previously described experiments would be consistent with any number of bases larger than two. The argument that the number is, in fact, three, or some multiple of three was shown by experiments which formed triple mutants composed of three (+) mutations or three (-) mutations. The result, and its interpretation depends to a large degree on the fact that the amino acid sequence coded by the region of the rIIB cistron under consideration seems not to be very crucial to the function of the whole rIIB protein. That seems clear from two facts: One is that the sequence of amino acids in that region can, for short segments, at least, be altered without loss of function judging from the mutual suppression of (+) and (-) mutations. A second fact is that the mutation r1589 discovered by CHAMPE and BENZER (1962) is an alteration which deletes much of the A cistron, but also part of the B cistron, and precisely that part of the B cistron, in which FC-0 maps. Even though that part of it is missing, the B cistron can function. It therefore seemed possible that the addition or deletion of an amino acid from that portion of the protein might not inactivate the function. If the coding ratio is three bases per amino acid, the combination of three (+) mutations or three (-) mutations would simply change the amino acid sequence between the terminal mutations and make a net increase or decrease of one amino acid. Five triple (+) and one triple (-) mutants were constructed by recombination (taking care not to make an improper crossing of a barrier). Each of those triple mutants had a wild phenotype, in contrast to any combination of two (+) or two (-) mutants which always had the r phenotype. This was interpreted to mean that the coding ratio was three, or some multiple of three bases per amino acid.

It still seemed possible, however, that the original FC-0 mutation was due to a double (or triple) base addition or deletion. All of the first experiments were derived from FC-0, and therefore all suppressors which were selected because they corrected the phenotype of this mutant, as well as second and third order suppressors would have been double (or triple) changes of opposite sign. In that case the coding ratio would be six (or nine). That this is unlikely was shown by study of frame-shift mutations of independent origin, four induced by acridine yellow and two induced by hydrazine. Two of the acridine yellow mutations proved to be (+) and two (-), the former suppressing FC-9, a suppressor with the (-) sign from the FC-0 series. The two hydrazine mutations were both (-), suppressing FC-30 from the original series. If FC-0 was a double-base change, then the six independent mutants must also have resulted from double-base changes since they could bring about suppression of the FC-0 series. To maintain now that all of these mutations are double base changes would imply that both mutagens more or less regularly cause addition or deletion of adjacent base doublets, which seems rather improbable. Thus the coding ratio of three seems to be the most likely number.

At this point it seemed quite evident that the base sequence is read in multiples of three, and that a fixed starting point assures a proper register of the reading frame. An additional experiment increases the plausibility of the idea of fixed starting points. The rIIA and rIIB cistrons appear to be read completely independently of each other as shown by the cis-trans test, especially when frame shift mutations are employed in it. Thus, neither a (+) or a (-) frame shift mutation in rIIA affects the function of the B cistron, nor do frame shifts in B influence the function of the A cistron. The B cistron must, therefore, have its own starting point. Experiments with the deletion r1589 (referred to above) show that this starting point is not unique, but that another starting point can be utilized. The usual starting point of the B cistron must be absent not only because the deletion in r1589 appears to include the region between the A and B cistrons, but also because insertion of a frame-shift mutation in the

non-deleted section of A causes loss of the B function. Furthermore, when a second frame shift mutation of opposite sign is introduced in the segment of the A cistron which remains, the B cistron function is regained. The loss and regaining of B cistron function in r1589 strongly suggests that reading starts from a new point located in or ahead of the A cistron, and that the result is synthesis of a protein which can carry out B cistron function.

THE LYSOZYME STORY: A SPECIFIC TEST OF THE GENERAL NATURE OF THE GENETIC CODE.

From experiments of the types described, FRANCIS CRICK and his collaborators were able to amass evidence for the most important general features of the genetic code. Final proof that they, in fact, had made the correct interpretation of their experiments was achieved by close scrutiny of the lysozyme gene of phage T4. CRICK, BARNETT, BRENNER, and WATTS-TOBIN (1961) in discussing the future of the coding problem wrote "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,...". Several tests of this prediction by a large research team headed by GEORGE STREISINGER leaves no doubt as to the correctness of the theory, and is the final achievement which I will discuss on the role of bacteriophage T4 in clarifying our concept of the gene.

The rIIA and rIIB cistrons, for which the most complete genetic information and material were available, had one serious shortcoming for the final proof of the relationship between the gene and the protein for which it is responsible: for these genes the proteins were unknown. The fact that phage lysozyme (identified by KOCH and DREYER, 1958, in tails and lysates of phage T2) could be isolated in milligram amounts from T4 lysates led STREISINGER (1964) to study the gene responsible for that protein in phage T4. By use of frame-shift mutations in the structural gene for lysozyme, this group of investigators, which included STREISINGER, TSUGITA, TERZAGHI, OKADA, INOUYE, EMRICH, NEWTON, and AKABOSHI, showed that the gene does, in fact, use triplets which are read from a fixed starting point to code for individual amino acids. The prediction made by CRICK and his collaborators was thereby confirmed completely.

The procedure, while it involved many steps, is not complicated in principle. Acridine-induced lysozyme mutations were first classified into (+) and (-) frame-shift categories by mutual suppression experiments similar to those previously described for the rIIB system. The pseudo-wild lysozyme was then isolated from a double mutant phage carrying a (+) and a (-) mutation. This enzyme was compared with the lysozyme from a wild-type T4 lysate. The first comparison was made by use of column chromatography to separate the products of trypsin (and in some cases chymotrypsin) digestion. The wild-type lysozyme yielded 18 peptides in this procedure, and in one case, for example (TERZAGHI, OKADA, STREISINGER, EMRICH, INOUYE, and TSUGITA 1966), the lysozyme of the double mutant eJ42eJ44 was found to yield 17 peptides which were identical to 17 from the wild-type protein. Peptide no. 10, however, was absent, but a new peptide, designated no. 19, was found in the pseudo-wild digest. Analysis of the amino acid sequences of no. 10 and no. 19 showed that both contained 8 amino acids, and both had the -COOH terminal sequence -Ala-Ala-Lys. The sequence of amino acids preceding those three was different in the two peptides, however. The wild-type sequence was Ser-Pro-Ser-Leu-Asn- while in the pseudo-wild

case it was Val-His-His-Leu-Met-. Similar analysis on a chymotryptic digest again revealed a one-peptide difference. Comparison of the amino acid sequences of the distinguishing peptides here indicated a common -NH_2 terminal pair consisting of Thr-Leu-. These were in the two cases followed by the same two sequences found in the tryptic peptides, namely, -Ser-Pro-Ser-Leu-Asn in the wild type and -Val-His-His-Leu in the double mutant. The tryptic peptide showed that the sequence following the five-amino-acid difference was the same in the wild and pseudo-wild proteins and the chymotryptic peptide showed that the sequence preceding it was similar. In that way it verified the prediction made by CRICK et al. that a continuous sequence of amino acids, in this case five, should be changed in the double mutant, and not two as conventional theory would predict.

The final proof of the general nature of the code came from correlating the amino acid sequences found in the wild-type and the pseudo-wild polypeptides with the coding triplets established through *in vitro* studies largely by the research teams headed by M. NIRENBERG and by H. G. KHORANA (see the Cold Spring Harbor Symposium for Quantitative Biology, volume 31 [1966] for exhaustive references.) Using the triplets proposed by those investigators, a sequence of bases was written that would code for the amino-acid sequence in the wild-type peptide no. 10. With the deletion from this sequence of a single adenine at one point and addition to it of a single adenine or guanine fifteen bases farther along, a new set of triplets was generated which coded for the five amino acids found in peptide no. 19 from the pseudo-wild lysozyme.

That result hardly seemed explainable by chance. Nevertheless, the test has been extended to at least four other double- or triple-mutant combinations (STREISINGER, OKADA, EMRICH, NEWTON, TSUGITA, TERZAGHI, and INOUYE 1966): A second example of a single base addition corrected by a single base deletion was analysed. Two double-mutant cases carried a two-base addition together with a single base addition resulting in the insertion of an extra amino acid in the polypeptide as well as an altered sequence between the mutations. The fifth case was a triple mutant which contained one two-base addition and two single-base deletions. Each pseudo-wild peptide was shown to be a possible transformation from the wild-type sequence resulting from single- or double-base changes and consistent with the coding triplets determined from the *in vitro* experiments.

The data thus provided by analysis of frame-shift mutations, first in the genetic and then in the chemical sense, seem to have established beyond a reasonable doubt that the information carried by a gene is written in a code in which sequential nucleotide triplets specify a sequence of amino acids. The final translation of the nucleotide message is dependent on initiating the reading process from a fixed starting point which is in proper register with the triplets.

In closing it may be recalled that STADLER's concept of the gene only fifteen years ago had little or no detail for want of adequate experimental material and methods for characterizing and dissecting it. Since that time studies with viruses and bacteria, notably bacteriophage T4 and *Escherichia coli*, have overcome the earlier deficiencies so that today we have a much clearer picture of the structure of the gene and the manner in which the structure is related to function. The resolving power of the microbial methods has even reached the molecular level. It must be kept in mind, however, that only the first details are available, and many others remain to be clarified. The signals for initiating and for terminating replication, transcription, and translation are poorly understood or unknown. And, although the degree of universality of the code already demonstrated gives the molecular biologist confidence that, in general, the same rules will apply to genes in eucaryotic systems, it would doubtless be foolish to anticipate no significant variations. It would be surprising to me if, fifteen years in the future, we were not to marvel

a little when we look back at the unsophisticated concept of the gene in 1969.

LITERATURE CITED

BENZER, S. 1955 Fine structure of a genetic region in bacteriophage. Proc. Nat. Acad. Sci. U. S. 41:344-354.

BENZER, S. 1957 The elementary units of heredity. In: The Chemical Basis of Heredity. Edited by W. D. McElroy and B. Glass. pp. 70-93. The Johns Hopkins Press, Baltimore.

BENZER, S. 1959 On the topology of the genetic fine structure. Proc. Nat. Acad. Sci. U. S. 45:1607-1620.

BENZER, S. 1961 On the topography of the genetic fine structure. Proc. Nat. Acad. Sci. U. S. 47:403-415.

BRENNER, S. 1957 On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. Proc. Nat. Acad. Sci. U. S. 43:687-694.

BRENNER, S., L. BARNETT, F. H. C. CRICK, and A. ORGEL 1961 The theory of mutagenesis. J. Mol. Biol. 3:121-124.

CHAMPE, S. P., and S. BENZER 1962 An active cistron fragment. J. Mol. Biol. 4:288-292.

CRICK, F. H. C., L. BARNETT, S. BRENNER, and R. J. WATTS-TOBIN 1961 General nature of the genetic code for proteins. Nature 192:1227-1232.

HERSHEY, A. D., and R. ROTMAN 1948 Linkage among genes controlling inhibition of lysis in a bacterial virus. Proc. Nat. Acad. Sci. U. S. 34:89-96.

KOCH, G., and W. J. DREYER 1958 Characterization of an enzyme of phage T2 as a lysozyme. Virology 6:291-293.

STADLER, L. J. 1954 The gene. Science 120:811-819.

STREISINGER, G. 1964 Gene-protein relationships. In International Symposium on Genes and Chromosomes, Structure and Function. National Cancer Inst. Monograph, No. 18, p. 1.

STREISINGER, G., Y. OKADA, J. EMRICH, J. NEWTON, A. TSUGITA, E. TERZAGHI, and M. INOUYE 1966 Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. 31:77-84.

TERZAGHI, E., Y. OKADA, G. STREISINGER, J. EMRICH, M. INOUYE, and A. TSUGITA 1966 Change of a sequence of amino acids in phage T4 lysozyme by acridine-induced mutation. Proc. Nat. Acad. Sci. U. S. 56:500-507.

TSUGITA, A. and H. FRAENKEL-CONRAT 1962 The composition of proteins of chemically evoked mutants of TMV RNA. J. Mol. Biol. 4:73-82.

WITTMANN, H. G. 1962 Proteinuntersuchungen an Mutanten des Tabakmosaikvirus als Beitrag zum Problem des genetischen Codes. Z. Vererbungslehre 93:491-530.