DNA SEQUENCING AND GENE STRUCTURE

Nobel lecture, 8 December, 1980

by

WALTER GILBERT

Harvard University, The Biological Laboratories, Cambridge, Massachusetts 02138, USA

When we work out the structure of DNA molecules, we examine the fundamental level that underlies all process in living cells. DNA is the information store that ultimately dictates the structure of every gene product, delineates every part of the organism. The order of the bases along DNA contains the complete set of instructions that make up the genetic inheritance. We do not know how to interprete those instructions; like a child, we can spell out the alphabet without understanding more than a few words on a page.

I came to the chemical DNA sequencing by accident. Since the middle sixties my work had focussed on the control of genes in bacteria, studying a specific gene product, a protein repressor made by the control gene for the *lac* operon (the cluster of genes that metabolize the sugar lactose. Benno Müller-Hill and I had isolated and characterized this molecule during the late sixties and demonstrated that this protein bound to bacterial DNA immediately at the beginning of the first gene of the three-gene cluster that this repressor controlled (1, 2). In the years since then, my laboratory had shown that this protein acted by preventing the RNA polymerase from copying the *lac* operon genes into RNA. I had used the fact that the *lac* repressor bound to DNA at a specific region, the operator, to isolate the DNA of this region by digesting all of the rest of the DNA with DNase to leave only a small fragment bound to the repressor, protected from the action of the enzyme. This isolated a twenty-five base-pair fragment of DNA out of the 3 million base pairs in the bacterial chromosome. In the early seventies, Allan Maxam and I worked out the sequence of this small fragment (3) by copying this DNA into short fragments of RNA and using on these RNA copies the sequencing methods that had been developed by Sanger and his colleagues in the late sixties. This was a laborious process that took several years. When a student, Nancy Maizels, then determined the sequence of the first 63 bases of the messenger RNA for the *lac* operon genes, we discovered that the *lac* repressor bound to DNA immediately after the start of the messenger RNA (4), in a region that lies under the RNA polymerase when it binds to DNA to initiate RNA synthesis. We continued to characterize the *lac* operator by sequencing a number of mutations (operator constitutive mutations) that damaged the ability of the repressor to bind to DNA. We wanted to determine more DNA sequence in the region to define the polymerase binding

site and other elements involved in *lac* gene control; however, that sequence was worked out in another laboratory by Dickson, Abelson, Barnes and Reznikoff (5). Thus by the middle seventies I knew all the sequences that I had been curious about, and my students (David Pribnow, and John Majors) and I were trying to answer questions about the interaction of the RNA polymerase and other control factors with DNA.

At this point, another line of experiments was opened up by a new suggestion. Andrei Mirzabekov came to visit me in early 1975. The purpose of his visit was twofold: to describe experiments that he had been doing using dimethyl sulfate to methylate the guanines and the adenines in DNA and to urge me to do a similar experiment with the *lac* repressor. Dimethyl sulfate methylates the guanines uniquely at the N7 position, which is exposed in major groove of the DNA double helix, while it methylates the adenines at the N3 position which is exposed in the minor groove (Fig. 1). Mirzabekov had used this property to attempt to determine the disposition of histones and of certain antibiotics on the DNA molecule by observing the blocking of the incorporation of radioactive methyl groups onto the guanines and adenines of bulk DNA. He urged me to use this groove specificity to learn something about the interaction of the *lac* repressor with the *lac* operator. However, the amounts of *lac* operator available were extremely small, and there was no obvious way of examining the protein sitting on DNA to ask which bases in the sequence the protein would protect against attack by the dimenthyl sulfate reagent.

It was not until after a second visit by Mirzabekov that an idea finally emerged. He and I and Allan Maxam and Jay Gralla had lunch together.

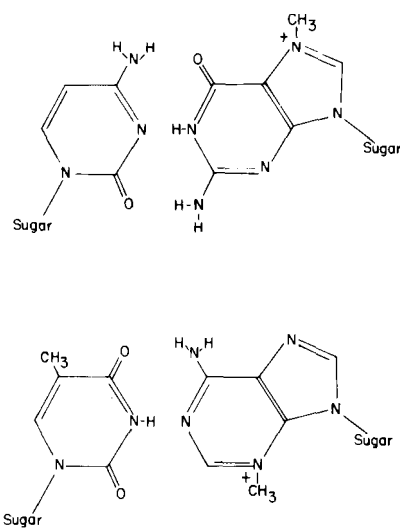


Figure 1. Methylated cytosine-guanine and thymine-adenine base pairs. The top of the figure shows cytosine-guanine base pair methylated at the N7 position of guanine. The bottom of the figure shows a thymine-adenine base pair methylated at the N3 position of adenine. The region above each of the base pairs is exposed in the major groove of DNA. The region below each of the base pairs lies between the sugar phosphate backbones in the minor groove of the DNA double helix.

During our conversation I had an idea for an experiment, which ultimately underlies our sequencing method. We knew we could obtain a defined DNA fragment, 55 base-pairs long, which carried near its center the region to which the *lac* repressor bound. This fragment was made by cutting the DNA sequentially with two different restriction engymes, each defining one end of the fragment (See fig. 2). Secondly, I knew that at every base along the DNA at which methylation occurred, that base could be removed by heat. Furthermore, once that had happened, only a sugar would be left holding the DNA chain together, and that sugar could be hydrolysed, in principle, in alkali to break the DNA chain. I put these ideas together by conjecturing that if we labelled one end of one strand of the DNA fragment with radioactive phosphate, we might determine the point of methylation by measuring the distance between the labelled end and the point of breakage. We could get such labelled DNA by isolating a DNA fragment (by length by electrophoresis through polyacrylamide gels) made by cutting with one restriction enzyme, labelling both ends of that fragment and then cutting it again with a second restriction enzyme to release two separable double-stranded fragments, each having a label at one end but not the other. Using polynucleotide kinase this procedure would introduce a radioactive label into the 5' end of one of the DNA strands of the fragment bearing the operator while leaving the other unlabelled (Fig. 2). If we then modified that DNA with dimethyl sulfate so that only an occasional adenine or guanine would be methylated, heated, and cleaved the DNA with alkali at the point of depurination, we would release among other fragments a labelled fragment extending from the unique point of labelling to the first point of breakage. Fig. 3 shows this idea. Any fragments from the other strand would be unlabelled, as would any fragments arising beyond the first point of breakage. If we could separate these fragments by size, as we could in principle by electrophoresis on a polyacrylamide gel, we might be able to associate the labelled fragments back to the known sequence and thus identify each guanine and adenine in the operator that had been modified by dimethyl sulfate. If we could do the modification in the presence of the *lac* repressor protein bound to the DNA fragment, then if the repressor lay close to the N7 of a guanine, we

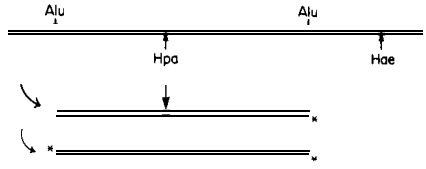


Figure 2. Procedure for obtaining a double-stranded DNA fragment uniquely labeled at one end of one strand. The figure shows the restriction cuts for the enzymes AluI (target AG/CT) and Hpa (target C/CGG producing an uneven end) in the neighborhood of the *lac* operator. The *lac* repressor is shown bound to the DNA. By cutting the DNA from this region first with the enzyme Alu, then labeling with radioactive P^{32} the 5' termini of both strands of the DNA with polynucleotide kinase, and then cutting in turn with the enzyme Hpa, we can isolate a DNA fragment that carries the binding site for the repessor uniquely labeled at one end of one strand.

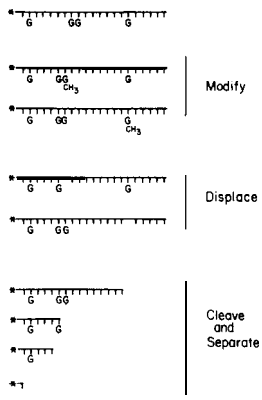


Figure 3. Outline of procedure to produce fragments of DNA by breaking the DNA at guanines. Consider an end-labeled strand of DNA. We modify an occasional guanine by methylation with dimethylsulfate. Heating the DNA will then displace that guanine from the DNA strand, leaving behind the bare sugar; cleaving the DNA with alkali will break the DNA at the missing guanine; the fragments are the separated by size, the actual size of the fragment (followed because it carries the radioactive label) determines the position of the modified guanine.

would not modify the DNA at that base, and the corresponding fragment would not appear in the analytical pattern.

I set out to do this experiment. Allan Maxam made the labelled DNA fragments, and I began to learn how to modify and to break the DNA. This involved analysing the release of the bases from DNA and the breakage steps separately. Finally the experiment was put together. Figure 4 shows the results: an autoradiogram of the electrophoretic pattern displays a series of bands extending downward in size from the full length fragment, each caused by the cleavage of the DNA at an adenine or a guanine. The same treatment of the DNA fragment with dimethyl sulfate, now carried out in the presence of the repressor produced a similar pattern, except that some of the bands were missing (lane one versus lane two in Figure 4). The experiment was clearly a success in that the presence of the repressor blocked the attack by dimethyl sulfate on some of the guanines and some of the adenines in the operator (6). I hoped that the size discrimination would be accurate enough to permit the assignment of each band in the pattern to a specific base in the sequence. This proved true because the spacing in the pattern, and the presence of light and dark bands, the dark bands corresponding to guanines and the light ones to adenines, were sufficiently characteristic to correlate the two. The guanines react about five times more rapdly with dimethyl sulfate while the methylated adenines are relesed from DNA more rapidly than the guanines during heating; the shift in intensities as a function of the time of heat treatment could by used to establish unambiguously which base was which. Furthermore, the gel pattern is so clear, that bands corresponding to fragments differing by one base were resolved. At this point it was evident that this technique could determine the adenines and guanines along DNA for distances of the order of 40 nucleo-

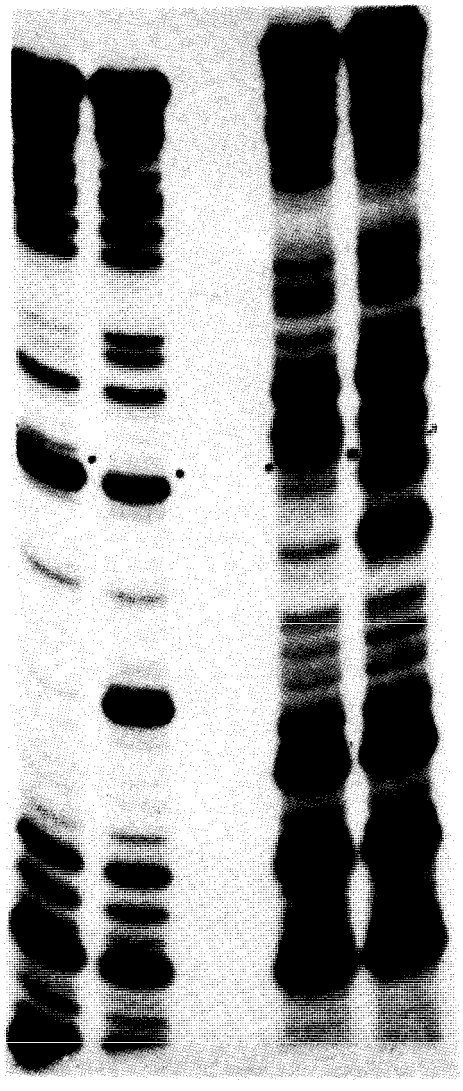
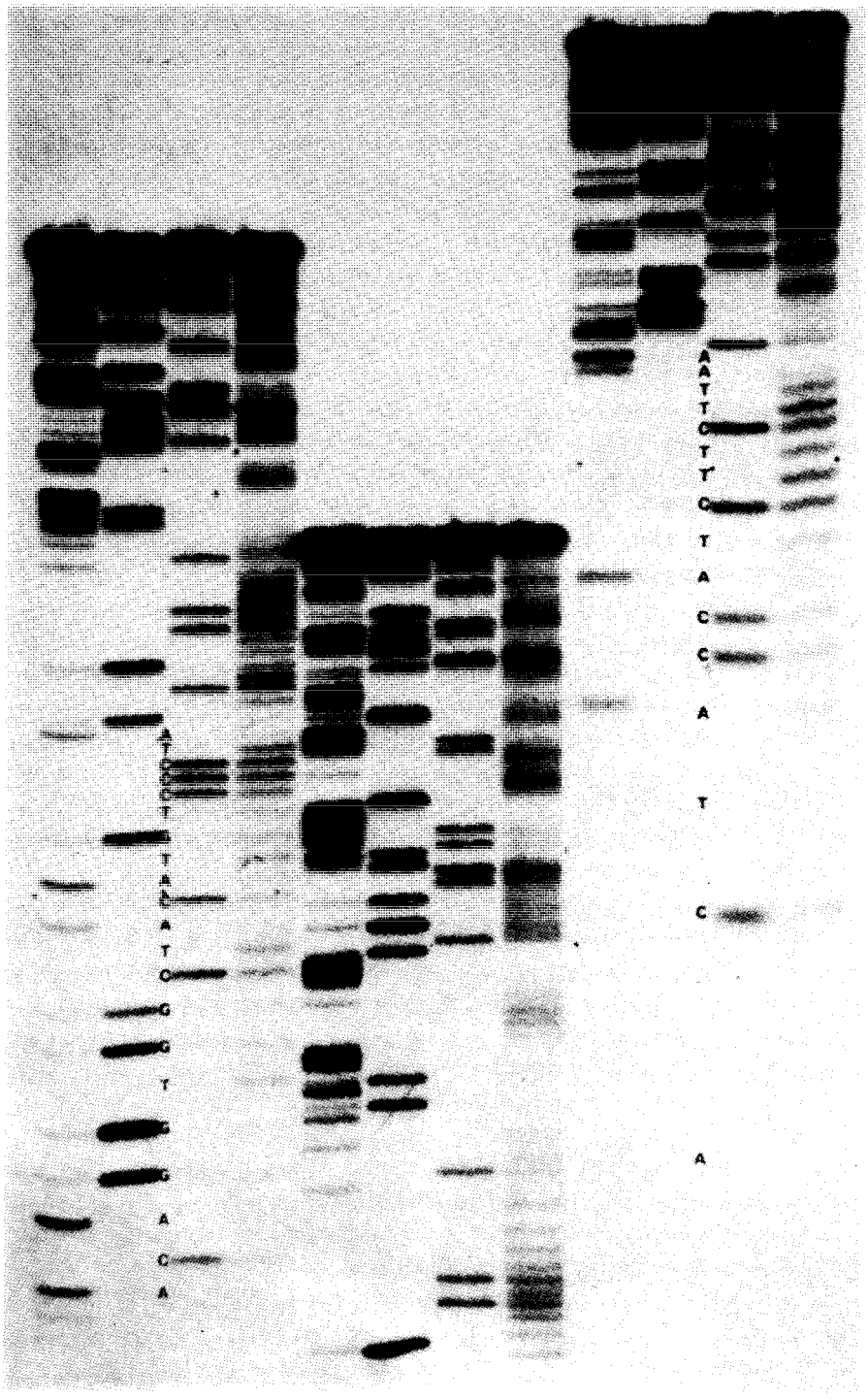


Figure 4. Methylation protection experiment with the *lac* repressor. The columns show the pattern of clevage along each strand of the 53-55 base long fragment bearing the *lac* operator. The second column from the left represents the DNA (labeled at the 5' end of the 53 base long strand) treated by dimethylsulfate, cleaved by heat and alkali. The dark bands correspond to breaks at guanines; the light bands are breaks at adenines. The second column of the figure reads from the bottom:
-G-GAAA--G--A---G---A-AA----A-A-AA-A-A-…
The first column shows this same double stranded piece of DNA treated with dimethylsulfate in the presence of the *lac* repressor. The repressor prevents the interaction of dimethytsulfate with the guanine a third of the way up the pattern and blocks the reaction with two adenines in the upper third of the pattern. The bands correspond to these two adenines represent fragments that differ in length by one base, 30 and 31 long. The right hand side of this pattern shows the same experiment done with the label at the end of the other DNA strand. The sequence at the far right would be:
-G-G-GGAA--G-GAG-GGA-AA-AA---...

tides. By determining the purines on one strand and the purines on the complementary strand (as Fig 4 does) one has in principle a complete sequencing method.

Having in hand a reaction that will determine and distinguish adenines and guanines, could we find reactions that would distinguish cytosines and thymines? Allan Maxam and I turned our attention to this end. (First we examined a second binding site for the *lac* repressor that lies a few hundred bases further along the DNA, under the first gene of the operon. This binding site has no physiological function. We could locate this binding site on a restriction fragment by repeating the methylation-protection experiment and identifying bases protected by the *lac* repressor. I used the methylation pattern to attempt to predict the positions of the adenines and the guanines in the unknown DNA sequence; Allan Maxam then used the wandering spot sequencing method of Sanger and his coworkers to determine the DNA sequence of this region to verify that we had made a successful prediction.) Allan Maxam then went on to do the next part of the development. We knew that hydrazine would attack the cytosines and thymines in DNA and damage them sufficiently, or eliminate them to form a hydrazone, so that a further treatment of the DNA with benzaldehyde followed by alkali, (or a treatment with an amine), would cleave at the damaged base. This soon gave us a similar pattern, but broke the DNA at the cytosines and thymines without discrimination. Allan Maxam then discovered that salt, one molar salt in the 15 molar hydrazine, altered the reaction to suppress the reactivity of the thymines. The two reactions together then positioned and distinguished the thymines and the cytosines in a DNA sequence. This last discrimination conceptually completed the method. To improve the discrimination between the purines, and to provide redundant information which would serve to make the sequencing more secure against errors, we used the fact that the methylated adenosines depurinate more rapidly, especially in acid, to release the adenosines preferentially and thus to obtain four reactions: one for A's, one preferential for G's, one for C's and T's, and one for C's determining the T's by difference. This stage of the work was completed within a few months. As the range of resolution on the gels was extended toward 100 bases, the cleavage at the pyrimidines was not satisfactory, the result of the incomplete cleavage was that the longer fragments contained a variety of internal damages and the pattern blurted out. After many months of searching an answer was found. A primary amine, aniline, will displace the hydrazine products and produce a beta elimination that releases the phosphate from the 3' position on the sugar, but it will not release the other phosphate, and the mobility of a DNA fragment with a blocked 3' phosphate bearing a sugar-aniline residue is different from the free phosphate ended chains from the other reactions. A secondary amine, piperidine, is far more effective and triggers both beta eliminations as well as eliminating all the breakdown products of the hydrazine reaction from the sugar. This reagent completed the DNA sequencing techniques (7). Although the development of the techniques continued for another nine months, they were distributed freely to other groups that wanted to use them. Fig. 5 shows an actual sequencing


A
A
T
T
C
T
T
C
T
A
C
C
A
T
C
A
T
C
G
T
A
C
A

pattern from the 1978 period, used in the work described in (8). Fig. 6 shows two examples of the chemistry (9).

The logic behind the chemical method is to divide the attack into two steps. In the first we use a reagent that carries the specificity, but we limit the extent of that reaction - to only one base out of several hundred possible targets in each DNA fragment. This permits the reaction to be used in the domain of greatest specificity: only the very initial stages of a chemical reaction are involved. The second step, the cleavage of the DNA strand, must be complete. Since the target has already been distinguished from the other bases along the DNA chain by the preliminary damage, we can use vigorous, quantitative reaction conditions. The result is a clean break, releasing a fragment without hidden damages, which is required if the mobilities of the fragments are to be very closely correlated so that the bands will not blur. (The specificity need be only about a factor of ten for the sequence to be read unambiguously.)

Today, later developments of the technique (9) have modified the guanine reaction and replaced the dimethyl sulfate adenosine reaction with a direct depurination reaction that releases both the adenines and guanines equally. These changes, and the introduction of the very thin gels by Sanger's group (10), now make it possible to read sequences out between 200-400 bases from the point of labelling. The actual chemical workup, the analysis on gels, and the autoradiography is the short part of the process. The major time spent in DNA sequencing is spent in the preparation of the DNA fragments and on the elements of stratagy. The speed of the sequencing comes only in part from the ability to read off quickly several hundred bases of DNA - at a glance. The more important element is the linear presentation of the problem. Rather than sequence randomly, one can begin at one end of a restriction map and move rationally through a gene - or construct the restriction map as one goes.

The first long sequence was done by a graduate student, Phillip Farabaugh, who used the new techniques to sequence the gene for the *lac* repressor (11). The protein sequence of this gene product had been worked out in the early seventies by Beyreuther and his coworkers (12). Since the amino-acid sequence was known, he could quickly (a few months) establish the DNA sequence. However the DNA sequence showed that there were errors in the protein sequence, two amino acids dropped at one place and eleven at another. Since

---

Figure 5. Actual sequencing pattern from the 1978 period. Products of four different chemical reactions, applied to a DNA fragment about 150 bases long, are electrophoresed on a polyacrylamide gel; three loadings produce sets of patterns that have moved different distances down the gel. The four columns correspond to reactions that break the DNA: I) primarily at the adenines, 2) only at the guanines, 3) at the cytosines but not the thymines, and 4) at both the cytosines and thymines. The very shortest fragments are at the bottom right hand side of the picture and the sequence is read up the gel recognizing first the band in the left hand column corresponding to A, a band in the two! right hand columns corresponding to a C, a band in the far right hand column corresponding to T, a band in the left hand column corresponding to A and so forth. After reading up as far aspossible, the sequence continues in the sets of bands at the left hand side of the gel and then still further in the pattern in the center of the gel. From the original photograph the sequence of the entire fragment can be read. The fragment is from the genomic DNA corresponding to the variable region of the lambda light chain of mouse immunoglobulin (8).

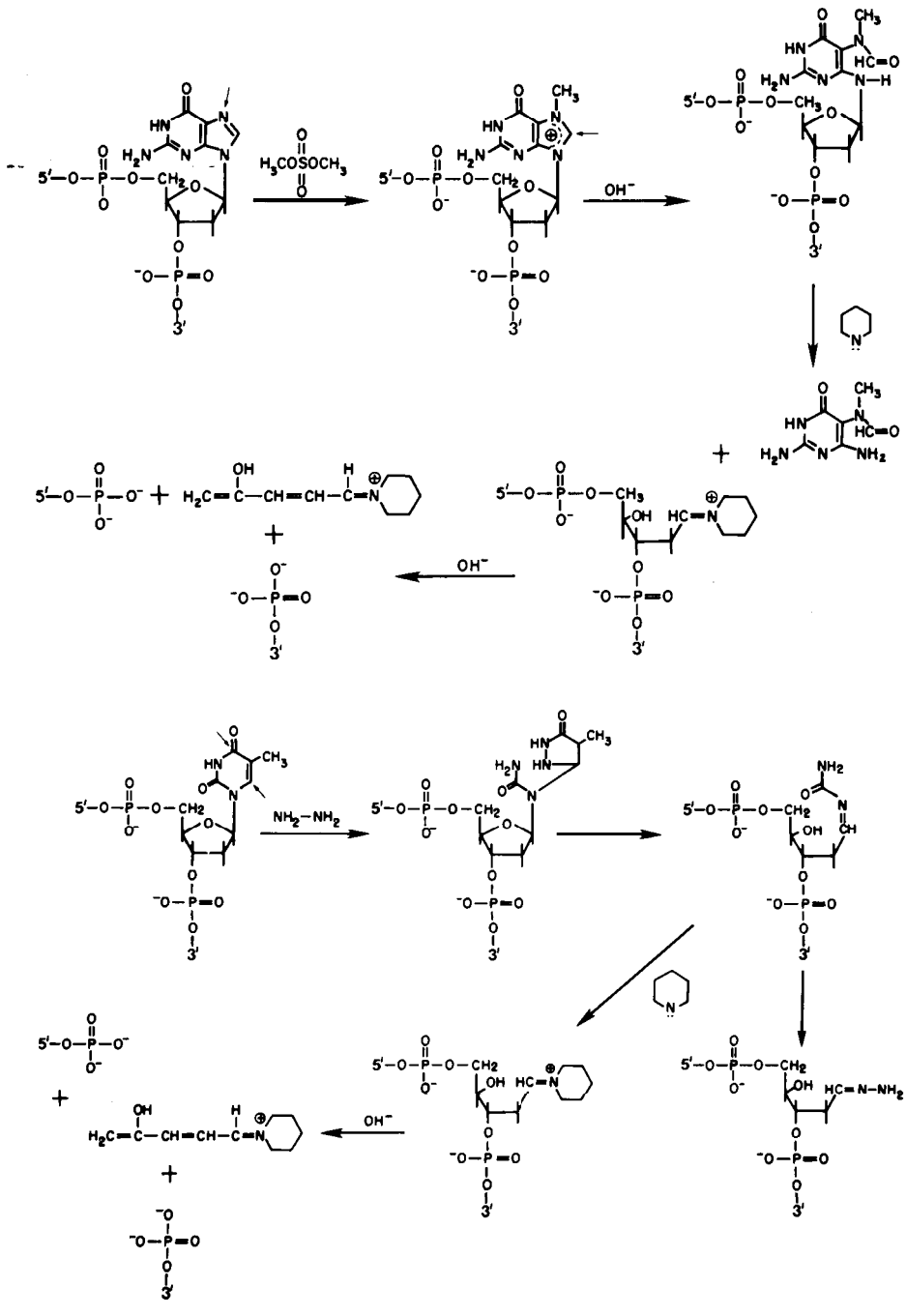


Figure 6. Examples of the detailed chemistry involved in breaking the DNA. Figure 6a. above, shows the guanine breakage. The guanines are first methylated with dimethylsulfate. The imidazole ring is opened by treatment with alkali (during the piperidine treatment). Piperidine displaces the base and then triggers two beta eliminations that release both phosphates from the sugar and cleave the DNA strand leaving a 3' and a 5' phosphate. Figure 6b. below, shows the hydrazinc attack on a thymine that breaks the DNA at the pyrimidines.

the protein sequence contains 360 amino acids, he had to work out a gene of 1080 bases. DNA sequencing is faster and more accurate than protein sequencing. The reason for this is that DNA is a linear information store. Because the chemistry of each restriction fragment is like any other, they differ only in length, there is no particular reason for loosing track of them, except for the very smallest. By sequencing across the joins between the fragments, one established an unambiguous order. Proteins, on the other hand, are strings of amino acids used by Nature to create a wide variety of chemistries. When a protein is fragmented, the fragments can exhibit quite different properties, some of which may be unusually unfortunate in terms of solubility or loss. There is no simple way of keeping account of the total content of amino acids, or of the order of fragments, as there is for DNA, where the length of the restriction fragments can easily be measured.

Jeffrey Miller and his coworkers had done an extensive analysis of the appearance of mutations in the *lac* repressor gene. Three sites in the gene are hotspots, at which the mutation rate is some 10 times higher than at other sites. DNA sequencing showed that at each of these sites there was a modified base, a 5-methyl cytosine, in the sequence (13). (The chemical sequencing detects the presence of the 5-methyl cytosine directly, because the methyl group suppresses completely the reactivity of this base in the hydrazine reaction. A blank space appears in the sequence, but on the other strand is a guanine.) The high mutation rate is a transition to a thymine. 5-methyl cytosine occurs at a low frequency in DNA, this observation shows that it is a mutagen. What is the explaination? Deamination of cytosine to uracil occurs naturally. If this occurred in DNA it could lead to a transition; however it usually does not, since there is an enzyme that scans DNA examining it for deoxyuridine (14). When it finds this base in DNA, mismatched or not, it breaks the glycosidic bond and removes the uracil. This is then recognized as a defect in DNA, and another group of enzymes then repair the depyrimidinated spot. However, 5-methyl cytosine deaminates to thymine - a natural component of DNA. On repair or resynthesis a transition will ensue. This whole argument explains why thymine is used in DNA - the extra methyl group serves to suppress the effects of the natural rate of deamination.

To find out how easy and how accurate DNA sequencing was, I asked a student, Gregor Sutcliffe, to sequence the ampicillin resistance gene, the beta-lactamase gene, of *E. coli.* This gene is carried on a variety of plasmids, including a small constructed plasmid, pBR322, in *E. coli.* All that he knew about the protein was an approximate molecular weight, and that a certain restriction cut on the plasmid inactivated that gene. He had no previous experience with DNA sequencing when he set out to work out the structure of DNA for this gene. After seven months he had worked out about 1000 bases of double-stranded DNA, sequencing one strand and then sequencing the other for confirmation. The unique long reading frame determined the sequence of the protein product of this gene, a protein of 286 residues (15). We thought that the DNA sequence was unambiguous. Luckily there was available, from Ambler's laboratory, partial sequence information about the protein which had

been obtained as a result of several years work attempting to develop a sequence for the beta-lactamase (16). This information, while not sufficient to determine the protein sequence directly, was adequate to confirm that the prediction of the DNA sequencing was correct. Sutcliffe then became very enthusiastic and sequenced the rest of the plasmid pBR322 during the next six months, to finish his thesis. He sequenced both strands of this 4362 base-pair long plasmid in order to confirm the sequence (17). The chemical sequencing is unambiguous, except for an occasional characteristic feature in the DNA fragment itself that causes it to move anomalously during the gel electrophoresis. As longer and longer strands are being analysed on the gel, a hairpin loop can form at one end of the fragment if the sequence is sufficiently self-complementary. As the fragmentation passes through this portion of the molecule, the mobilities on the gel do not decrease uniformly as a function of length, but some of the molecules move abberantly, a feature called compression, because the bands on an autoradiograph become close together, or can overlap to conceal one or more bases. This rare feature occurs about once every thousand bases. It is resolved by sequencing the opposite strand in the other direction along the double stranded molecule (or the same strand in the opposite chemical direction) because the hairpin will form when a different region of the sequence is exposed and the compression feature will occur in a different place in the sequence. If both strands of the DNA helix are sequenced, the sequence can be unambiguous.

THE STRUCTURE OF GENES

The first genes to be sequenced, those in bacteria, yielded an expected structure: a contiguous series of codons lying upon the DNA between an initiation signal and one of the terminator signals. Before the position at which the RNA copy will start, there lies a site for the RNA polymerase, interacting with the Pribnow box, a region of sequence homology lying one turn of the helix before the initial base of the messenger RNA, and also with another region of homology, thirty-five bases before the start. Thus one could understand the bacterial gene in terms of a binding site for the RNA polymerase, and further binding sites for repressors and activator proteins around and under the polymerase. Alternatively, the control on transcription could be exercised by a control of the termination function: new proteins or an elegant translation control (18) could determine whether or not the polymerase would read past a stop signal into a new gene.

When the first genes from vertebrates were transferred into bacteria by the recombinant DNA techniques and sequenced an entirely different structure emerged. The coding sequences for globin (19,20), for immunoglobulin (21), and for ovalbumin (22) did not lie on the DNA as a continuous series of codons but rather were interrupted by long stretches of non-coding DNA. The discovery of RNA splicing in adenovirus by Sharp and his coworkers (23) and Broker and Roberts and their coworkers (24) paved the way for this new structure.

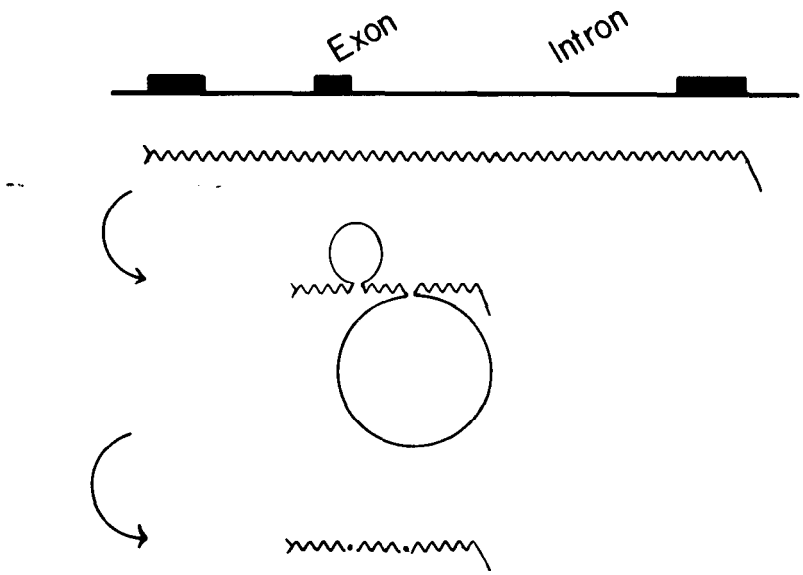


Figure 7. A transcription unit corresponding to alternating exons and introns. The whole gene, a transcription unit, is copied into RNA terminating in a poly (A) tail. The regions corresponding to introns are spliced out leaving a messenger RNA made up of the three exons, the regions that are expressed in the mature message.

They had shown that after the original transcription of DNA into a long RNA, regions of this RNA are spliced out: some stretches excised and the remaining portions fused together by an as yet undefined enzymatic process. The exons (25), regions of the DNA that will be expressed in mature message, are separated from each other by introns, regions of DNA that lie within the genetic element but whose transcripts will be spliced out of the message. Figure 7 shows this process: the original transcript of a gene (now thought of as a transcription unit) will undergo a series of splices before being able to function as a mature message in the cytoplasm. Figure 8 shows a few examples. Vertebrate genes can have many, eight, fifteen, even 50 exons (29,30), and the exons are for the most part short coding stretches separated by hundreds to several thousands of base pairs of intron DNA. The rapid sequencing has meant that we can work out the DNA sequence of any of these complex gene structures. But can we understand them?

The emerging generalization is that procaryotic genes have contiguous coding sequences while the genes for the highest eucaryotes are characterized by a complex exon-intron structure. As we move up from procaryotes, the simplest eucaryotes, such as yeasts, have few introns; further up the evolutionary ladder the genes are more broken up. (Yeast mitochondria have introns, are they an exception to this pattern?) Are we seeing the emergence of the intron-exon structure rising to ever greater degress of complexity as we move up to the vertebrates, or the loss of preexisting intron-exon structures as we move down to the simplest invertebrates and the procaryotes? One view considers the

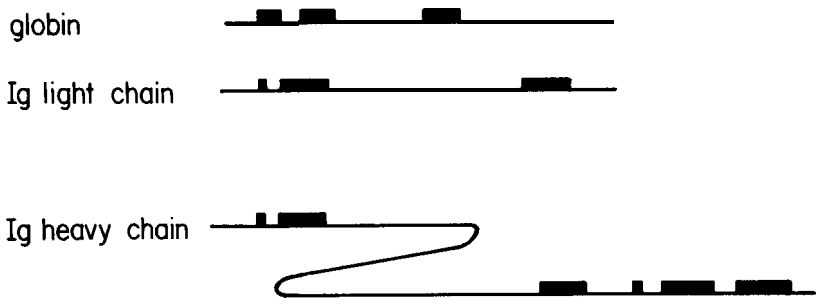


Figure 8. Examples of the intron-exon structure of a few genes. (1) The gene for globin is broken up by two introns into three exons (20). (2) The functional gene in a myeloma cell for the immunoglobulin lambda light chain is broken up into a short exon corresponding to the hydrophobic leader sequence, an exon corresponding to the V region, and then, after an intron of some thousand bases, an exon corresponding to the 112 amino acids of the constant region (26). (3) A typical gene for a gamma heavy chain of immunoglobulin (27, 28). The mature gene corresponds to a hydrophobic leader sequence, an exon corresponding to the variable region, and then, after a long intron. a series of exons; the first corresponding to the first domain of the constant region, the second corresponding to a 15 amino acid hinge region, the third corresponding to the second domain of the constant region, and the fourth exon corresponding to the third domain of the constant region.

splicing as an adaptation that becomes ever more necessary in more highly structured organisms. The other view considers the splicing as lost if the organism makes a choice to simplify and to replicate its DNA more rapidly, to go through more generations in a short time, and thus to be under a significant pressure to restrict its DNA content (31).

What role can this general intron-exon structure play in the genes of the higher organisms? Although most genes that have been studied have this structure, there are two notable exceptions: the genes for the histones and those for the interferons. This last demonstrates that there can be no absolutely essential role that the introns must play, there can be no absolute need for splicing in order to express a protein in mammalian cells. Although there is a line of experiments that shows that some messengers must have at least a single splice made before they can be expressed, there is no evidence that the great multiplicity of splices are needed. There is a pair of genes for insulin the rat, that differ in the number of introns; both are expressed - which demonstrates that the intron that splits the coding region of one of them, has no essential role, in cis, in the expression of that gene (32). Although a common conjecture is that the splicing might have a regulatory role, so far there is no tissue dependent splicing pattern that could be interpreted as showing the existance of a gene (or tissue) specific splicing enzyme.

The introns are much longer than the exons. Their DNA sequence drifts rapidly by point mutation and small additions and deletions (accumulating changes as rapidly as possible, at the same rate as the silent changes in codons). This suggests that it is not their sequence that is relevant, but their length. Their function is to move the exons apart along the chromosome.

A consequence of the separation of exons by long introns is that the recombination frequency, both illegitimate and legitimate, between exons will be higher (25). This will increase the rate, over evolutionary time, at which the exons, representing parts of the protein structures, will be shuffled and reassorted to make new combinations. Consider the process by which a structural domain is duplicated to make the two domain structure of the light chain of the immunoglobulins (or duplicated again to make the four domain structure of the heavy chain, or combined to make the triple structure represented by ovomucoid (29)). Classically, this involved a precise unequal crossing over that fused the two copies of the original gene, in phase, to make a double length gene. As Figure 9 shows, this process involves an extremely rare, precise illegitimate event (a recombination event that leads to the fusing of two DNA sequences at a point where there is no matching of sequence) that has as its consequence the synthesis at a high level of the new, presumably more useful, double length gene product. Consider the same process against the background of a general splicing mechanism. Again, the process of forming the double gene must involve an illegitimate recombination event, but now that event can occur anywhere within a stretch of 1000 to 10,000 bases flanking the 3' side of one copy and the 5' side of the other to form an intron separating the genes for the two domains. From a long transcript across this region, even inefficient splicing may produce the new double-length gene product. This will happen some 10^6 to 10^8 times more rapidly than the classical process because of the many

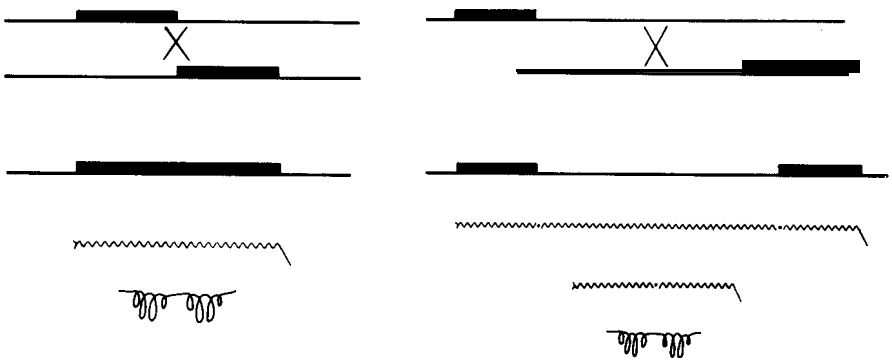


Figure 9. A double-length gene product arises through unequal crossing over. On the left, figure 9a, is the classical process by which a gene corresponding to a single polypeptide chain might have its length doubled by a crossing over. The top two lines indicate two coding regions, brought into accidental apposition by some act of illegitimate recombination which fuses the carboxy terminal region (the 3' end) of one copy of the gene to the amino terminal region (the 5' end) of the other. This rare illegitimate event (involving no sequence matching) would, if it occurs in phase. produce a double-length gene which could code for a double-length RNA which in turn translates into a double-length protein containing the reiteration of a basic domain. Figure 9b on the right illustrates the same process occurring in the presence of the splicing function. Now the unequal crossing over can occur anywhere to the *3'* side of one copy of the gene and anywhere in front of the 5' end of the other copy of the gene to produce a gene containing two exons separated by long intron. I conjecture that the long transcript of this region now will be spliced at some low frequency to produce a mature message encoding the reiterated protein.

different combinations of sites at which the recombination can occur. If the long transcript can be spliced, even at a low frequency, some of the double-length product can be made. This is a faster way for evolution to form the final gene: proceeding through a rapid step to a structure that can produce a small amount of the useful gene product. Small mutational steps can be selected to produce better splicing signals and thus more of the gene product. If the splicing signals already exist, recombination within introns provides an immediate way to build polymeric structures out of simpler units. One would predict that polymeric structures, made up of simpler units, will be found to have genes in which the intron-exon structure of the primitive unit is repeated, separated again by introns. That is the case.

The rate of legitimate recombination between the exons of a gene will be increased by the introns. Consider two mutations to better functioning, arising in different parts of a gene and spreading, by selection, through the population. Classically, both mutations could end up in a single polypeptide chain, after both genes find their way into a single diploid individual, by homologous recombination within the gene. Figure 10 shows that this process also should be speeded some 10 to 100 fold by spreading the exons apart. This effect will be strongest if the exons can evolve separately - if they represent structures that can accumulate successful changes independantly.

Furthermore, one can change the pattern of exons by changing the initiation or termination of the RNA transcript, to add extra exons or to tie together exons from one region of the DNA to exons from another. This has been observed in adenovirus, and is found in notable examples in the immunoglobulins in which exons can be added or subtracted to the carboxy terminus of the heavy chain to modify the protein. Hood's laboratory has shown that this process is used to switch between two different forms of an IgM heavy chain (33). A membrane bound form is synthesized by a longer transcript, which splices on two additional exons and splices out part of the last exon of the shorter transcript. The shorter transcript synthesises a secreted form of the

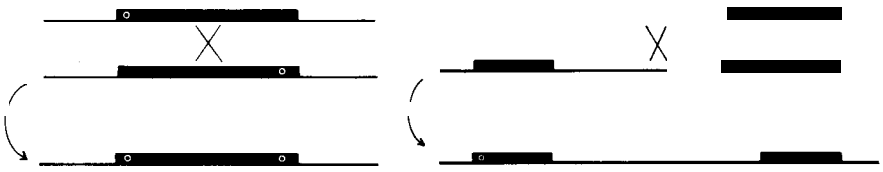


Figure 10. Introns speed legitimate recombination. Figure 10a, on the left, shows the classical pattern by which two mutations, one occurring in one copy of the gene, at the left end, and the other occurring in the other copy of the gene, at the right hand end, might get together by recombination happening in the homologous stretch of DNA that separates the two mutations. This recombination can create a single gene carrying both mutations. On the right, figure 10b, the same process happening in a gene in which the mutations occur in separate exons separated by an intron. Now the recombination can occur anywhere, either in the exon or within the intron, to produce a new gene carrying both mutations. Since the rate of recombination will be directly proportional to the distance along the DNA between the mutations, it will be faster.

protein. In a similar way the switch of the V region from IgM to an IgD constant region is probably the result of a different, still longer, transcript which splices across to attach the V region exons to the new constant region exons of the delta class. These combinations of genes have certainly been created by recombination events within the DNA that ultimately becomes the intron of the longer transcription unit.

The most striking prediction of this evolutionary view is that separate elements defined by the exons have some functional significance, that these elements have been assorted and put together in new combinations to make up the proteins that we know. Gene products are assembled out of prevously achieved solutions of the structure-function problem. Clear examples of this are still meager. The hydrophobic leader sequence which is involved in the transfer of proteins through membranes, and which is trimmed off after the secretion, is often on separate exons-most notably in the immunoglobulins (see Figure 8), but also in ovomucoid (29). In the pair of genes for insulin in the rat, a product of recent duplication (32), the two chains of insulin lie on separate in exons one gene, on a single exon in the other. The ancestral gene (the common structure in other species (34)) has the additional intron-suggesting that the gene was put together originally from separate pieces. The gene for lysozyme is broken up into four exons; the second one carries the critical amino acids of the active site and most of the substrate contacts (35). In the gene for globin the central exon encodes almost all of the heme contacts. Figure 11 shows a schematic disection of the molecule. A recent experiment (36) has shown that the polypeptide that corresponds to the central exon in itself is a heme binding "miniglobin"; the side exons have provided polypeptide material to stablize the protein.

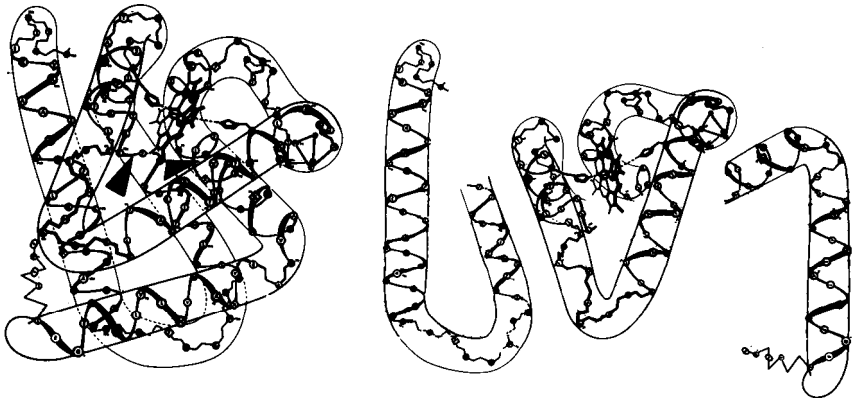


Figure 11, A schematic dissection of globin into the product of the separate exons. At the left the black arrows show the points at which the structure of a chain of globin is interrupted by the introns. (The structures of the chains of globin and of myoglobin are very similar, the schematic structure shown is myoglobin.) The introns interrupt the protein in the alpha-helical regions to break the protein into three portions shown on the right. The product of the central exon surrounds the heme; the products of the other two exons, I conjecture, wrap around and stablize the protein.

At the same moment that the rapid sequencing methods and the molecular cloning gave us the promise of being able to work out the structure of any gene, the ability to achieve a complete understanding of the genetic material, Nature revealed herself to be more complex than we had imagined. We can not read the gene product directly from the chromosome by DNA sequencing alone. We must appeal to the sequence of the actual protein, or at least the sequence of the mature messenger RNA, to learn the intron-exon structure of the gene. Nonetheless the hope exists, that as we look down on the sequence of DNA in the chromosome, we will not learn simply the primary structure of the gene products, but we will learn aspects of the functional structure of the proteins - put together over evolutionary time as exons linked through introns.

My interest in biology has always centered on two problems: how is the genetic information made manifest? and how is it controlled? We have learned much about the way in which a gene is translated into protein. The control of genes in prokaryotes is well understood, but for eukaryotes the critical mechanisms of control are still not known. The purpose of research is to explore the unknown. The desire for new knowledge calls forth the answers to new questions.

I owe a great debt to my students and collaborators over the years; the greatest to Jim Watson who stimulated my interest in molecular biology, to Benno Müller-Hill with whom I worked on the *lac* repressor, and to Allan Maxam with whom I developed the DNA sequencing.

BIBLIOGRAPHY

1 Gilbert, Walter and Müller-Hill, Benno "Isolation of the *Lac* Repressor" *Proc. Natl. Acad. Sci. USA.556,* 1891-1898 (1966).

2 Gilbert, Walter and Müller-Hill, Benno "The *Lac* Operator is DNA" *Proc. Natl. Acad. Sci. USA* 58, 2415-2421 (1967).

3 Gilbert, Walter and Maxam. Allan "The Nuclcotide Sequence of the *Lac* Operator" Proc. *Natl. Acad. Sci. USA 70,* 3581-3584 (1973).

4 Gilbert, W.. Maizels. N. and Maxam. A. "Sequences of Controlling Regions of the Lactose Operon" Cold Spring Harbor Symposium on Quantitative Biology 38. 845-855 (1973).

5 Dickson, R., Abelson, J., Barnes, W. and Reznikoff, W. "Genetic Regulation: the *Lac* Control Region" Science 187, 27-35 (1975).

6 Gilbert, Walter, Maxam, Allan and Mirzabekov, Andrei "Contacts Between the *lac* Repressor and DNA Revealed by Methylation" in Control of Ribosome Synthesis", 139-148, Alfred Benzon Symposium IX, Munksgaard 1976.

7 Maxam, Allan M. and Gilbert, Walter "A New Method for Sequencing DNA" *Proc. Natl. Acad. Sci. USA 74, 560-564* (1977).

8 Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. and Gilbert, W. "Sequence of a Mouse Germ-line Gene for a Variable Region of an Immunoglobulin Light Chain" *Proc. Natl. Acad. Sci. USA 75,* 1485-1489 (1978).

9 Maxam, Allan, M., and Gilbert, Walter "Sequencing End-Labelled DNA with Base-Specific Chemical Cleavages" *Methods In Enzmology 65,* 499-560 (editors K. Moldave and L. Grossman) (1980).

10 Sanger, F. and Coulson, A. R. "The Use of Thin Acrylamide Gels for DNA Sequencing" *FEBS Letters 87,* 107-110 (1978).

11 Farabaugh, Philip J. "Sequence of the *lacI* Gene" *Nature 274, 765-769* (1978).

12 Beyreuther, K., Adler, K., Fanning, E., Murray, C., Klemm, A. and Geisler, N. "Amino-Acid Sequence of *lac* Repressor from *Escherichia coli" Eur. J. Biochem. 59,* 491-509 (1975).

13 Coulondre, Christine, Miller, Jeffrey H., Farabaugh, Philp J. and Gilbert, Walter "Molecular Basis of Base Substitution Hotspots in *Escherichia coli" Nature 274, 775-780* (1978).

14 Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B. and Sperens, B. "DNA *N*- Glycosidases" *J. Biol. Chem. 252, 3286-3294* (1977).

15 Sutcliffe, J. Gregor "Nucleotide Sequence of the Ampicillin Resistance Gene *of Escherichia coli* Plasmid pBR322" *Proc . Natl. Acad. Sci. USA 75,* 3737-3741 (1978).

16 Ambler, R. P. and Scott, G. K. "The Partial Amino Acid Sequence of the Penicillinase Coded by the *Escherichia coli* Plasmid R6K" *Proc. Natl. Acad. Sci. USA 75, 3732-3736 (1978).*

17 Sutcliffe, J. G. "Complete Nucleotide Sequence of the *Escherichia coli* Plasmid pBR322" Cold Spring Harbor Symposium 43, 77-90 (1978).

18 For a review, see: Yanofsky, C. "Attenuation in the Control of Expression of Bacterial Opera*ons" Nature 289,* 751-758 (1981).

19 Tilghman, S. M., Tiemeister, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J. V. and Leder, P. "Intervening Sequence of DNA Identified in the Structural Portion of a Mouse Beta-Globin Gene" *Proc. Natl. Sci. USA 75, 725-729 (1978).*

20 Konkel, D. A., Tilghman, S. M. and Leder, P. "The Sequence of the Chromosomal Mouse Beta-Globin Major Gene" *Cell* 15, 1125-1132.

21 Brack, C. and Tonegawa, S. "Variable and Constant Parts of the Immunoglobulin Light Chain of a Mouse Myeloma Cell are 1250 Nontranslated Bases Apart" *Proc. Natl. Acad. Sci. USA 74,* 5652-5656 (1977).

22 Breathnach, R., Mandel, J. L. and Chambon, P. "Ovalbumin Gene is Split in Chicken DNA" *Nature 270,* 314-319 (1977).

23 Berget, Susan M., Moore, Claire and Sharp, Phillip A. "Spliced Segments at the 5' Terminus of Adenovirus 2 Late mRNA" *Proc. Natl. Acad. Sci. USA 74,* 3171-3175 (1977).

24 Chow, L. T., Gelinas; R. E., Broker, T. R. and Roberts, R. J. "An Amazing Sequence Arrangement at the 5' Ends of Adenovirus 2 Messenger RNA" Cell 12, l-8 (1977).

25 Gilbert, Walter "Why Genes in Pieces?" *Nature 271, 501 (1978).*

26 Bernard, O., Hozumi, N. and Tonegawa, S. "Sequence of Mouse Immunoglobulin Light Chain Genes Before and After Somatic Changes" *Cell 15*, 1133-1144 (1978).
27 Sakano, H., Rogers, J. H., Hüppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. and Tonegawa, S. "Domains and the Hinge Region of an Immunoglobulin Heavy Chain Are Encoded in Separate DNA Segments" *Nature* 277, 627-633 (1979).
28 Honjo, T., Obata, M., Yanawaki-Kataoka, Y., Kataoka, T., Kawakami, T., Takahashi, N. and Mano, Y. "Cloning and Complete Nucleotide Sequences of Mouse Gamma 1 Chain Gene" Cell *18*, 559-568 (1979).
29 Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R. and O'Malley, B. W. "Ovomucoid Intervening Sequences Specify Functional Domains and Generate Protein Polymorphisms" Cell *21*, 681-687 (1980).
30 Yamado, Y., Avvedimento, V. E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I., and de Crombrugghe, B. "The Collagen Gene: Evidence for its Evolutionary Assembly by Amplification of a DNA segment Containing an Exon of 54 bp." *Cell 22, 887-892* (1980).
31 Doolittle, W. F. "Genes in Pieces: Were They Ever Together?" *Nature 272*, 581 (1978).
32 Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R., and Tizard, R. "The Structure and Evolution of the Two Non-Allelic Rat Preproinsulin Genes" *Cell 18, 545-558* (1979).
33 Early, P., Rogers, F., Davis, M., Calami, K., Bond, M., Wall, R. and Hood, L. "Two mRNA's Can be Produced from a Single Immunoglobulin Gene by Alternative RNA Processing Pathways" *Cell 20, 313-319* (1980).
34 Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. "The Evolution of Genes: The Chicken Preproinsulin Gene" *Cell 20, 555-566* (1980).
35 Jung, Alexander, Sippel, Albrecht, E., Grez, Manuel and Schütz, Günther "Exons Encode Functional and Structural Units of Chicken Lysozyme *Proc. Natl. Acad. Sci. USA 77*, 5759-5763 (1980).
36 Craik, Charles, S., Buchman, Steven R. and Beychok, Sherman "Characterization of Globin Domains: Heme Binding to the Central Exon Product" *Proc. Natl. Acad. Sci. USA 77*, 1384-1388 (1980).

1466
SCIENTIFIC AMERICAN OFFPRINTS

Useful Proteins from Recombinant Bacteria

by Walter Gilbert and Lydia Villa-Komaroff

SCIENTIFIC AMERICAN
APRIL 1980
VOL. 242, NO. 4 PP. 74-94

PUBLISHED BY **W. H. FREEMAN AND COMPANY** 41 MADISON AVENUE, NEW YORK, NEW YORK 10010

Copyright © 1980 by Scientific American, Inc. All rights reserved. Printed in the U.S.A. No part of this offprint may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. The trademark and tradename "SCIENTIFIC AMERICAN" and the distinctive logotype pertaining thereto are the sole property of and are registered under the name of, and are used herein under license from, Scientific American, Inc.

Useful Proteins from Recombinant Bacteria

Bacteria into which nonbacterial genes have been introduced are able to manufacture nonbacterial proteins. Among the proteins made by recombinant-DNA methods are insulin and interferon

by Walter Gilbert and Lydia Villa-Komaroff

A living cell is a protein factory. It synthesizes the enzymes and other proteins that maintain its own integrity and physiological processes, and (in multicelled organisms) it often synthesizes and secretes other proteins that perform some specialized function contributing to the life of the organism as a whole. Different kinds of cells make different proteins, following instructions encoded in the DNA of their genes. Recent advances in molecular biology make it possible to alter those instructions in bacterial cells, thereby designing bacteria that can synthesize nonbacterial proteins. The bacteria are "recombinants." They contain, along with their own genes, part or all of a gene from a human cell or other animal cell. If the inserted gene is one for a protein with an important biomedical application, a culture of the recombinant bacteria, which can be grown easily and at low cost, will serve as an efficient factory for producing that protein.

Many laboratories in universities and in an emerging "applied genetics" industry are working to design bacteria able to synthesize such nonbacterial proteins. A growing tool kit of "genetic engineering" techniques makes it possible to isolate one of the million-odd genes of an animal cell, to fuse that gene with part of a bacterial gene and to insert the combination into bacteria. As those bacteria multiply they make millions of copies of their own genes and of the animal gene inserted among them. If the animal gene is fused to a bacterial gene in such a way that a bacterium can treat the gene as one of its own, the bacteria will produce the protein specified by the animal gene. New ways of rapidly and easily determining the exact sequence of the chemical groups that constitute a molecule of DNA make it possible to learn the detailed structure of such "cloned" genes. After the structure is known it can be manipulated to produce DNA structures that function more efficiently in the bacterial cell.

In this article we shall first describe some of these techniques in a general way and then tell how we and our colleagues Argiris Efstratiadis, Stephanie Broome, Peter Lomedico and Richard Tizard applied them in our laboratory at Harvard University to copy a rat gene that specifies the hormone insulin, to insert the gene into bacteria and to get the bacteria to manufacture a precursor of insulin. In an exciting application of this technology Charles Weissmann and his colleagues at the University of Zurich recently constructed bacteria that produce human interferon, a potentially useful antiviral protein.

DNA, RNA and Proteins

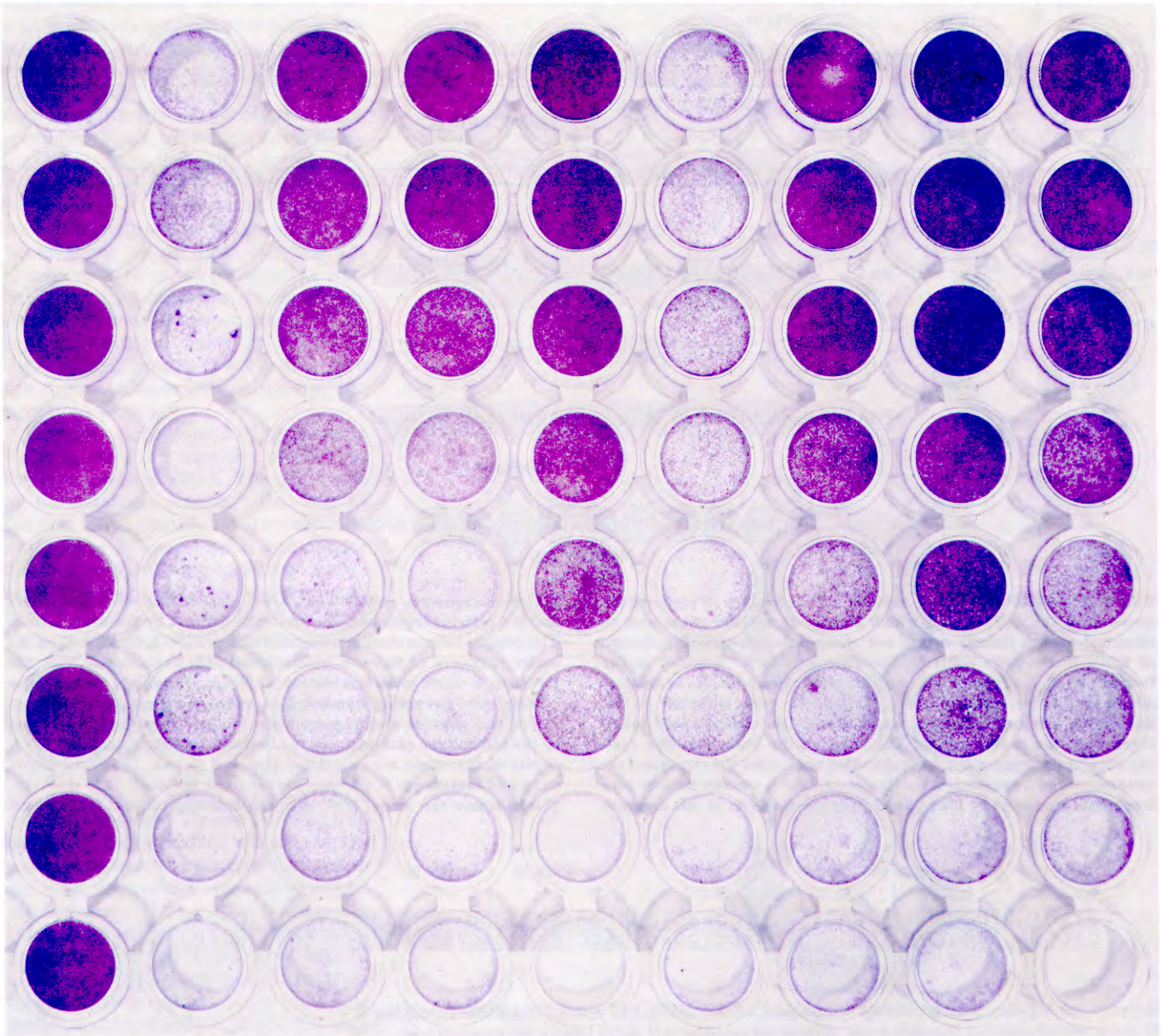
Cells make proteins by translating a set of commands arrayed along a strand of DNA. This hereditary information is held in the order of four chemical groups along the DNA: the bases adenine, thymine, guanine and cytosine. In sets of threes along DNA these bases specify which amino acids, the fundamental building blocks of proteins, are to be used in putting the protein together; the correspondence between specific base triplets and particular amino acids is called the genetic code. The part of a DNA molecule that incorporates the information to specify the structure of a protein is called a structural gene.

To act on this information the cell copies the sequence of bases from its genetic storehouse in DNA into another molecule: messenger RNA. A strand of DNA serves as a template for the assembly of a complementary strand of RNA according to base-pairing rules: adenine always pairs with uracil (which in RNA replaces DNA's thymine) and guanine pairs with cytosine. In animal cells transcription takes place in the nucleus of the cell. The messenger-RNA molecules carry the information out of the nucleus into the cytoplasm, where a complex molecular machine translates it into protein by linking together the appropriate amino acids. In bacteria, which have no nucleus, transcription and translation take place concurrently. The messenger RNA serves as a temporary set of instructions. Which proteins the cell makes depends on which messengers it contains at any given time; to make a different protein the cell makes a new messenger from the appropriate structural gene. The DNA in each cell contains all the information required at any time by any cell of the organism, but each cell "expresses," or translates into protein, only a specific small portion of that information. How does the cell know which structural genes to express?

Along with the structural information, a DNA molecule carries a series of regulatory commands, also written out as a sequence of bases. The simplest of these commands say in effect "Start here" or "Stop here" both for the transcription and for the translation steps. More complicated commands say when and in which type of cell a specific gene should be used. The genetic code is the same in all cell nuclei, a given structural sequence specifying the same protein in every organism, but the special commands are not the same in bacteria and in animal cells. One of the most surprising differences was discovered only in the past two years. The information for a bacterial protein is carried on a contiguous stretch of DNA, but in more complicated organisms, such as pigs and people, the structural information is broken up into segments, which are separated along the gene by long stretches of other DNA called intervening DNA or "introns." In such a cell a long region (often 10 times more than might be needed) is transcribed into RNA. The cell then processes this long RNA molecule, removing the sequence of bases that does not code for the protein and splicing together the rest to make a messenger-RNA molecule that carries essentially just the "start," the structural sequence and the "stop" needed for translation.

To persuade a bacterium to make a nonbacterial protein one must put into bacteria a DNA molecule that has a sequence of bases specifying the protein's amino acids as well as the bacterial commands for transcription and translation. Moreover, the inserted DNA must be treated by the bacterium as its own so that it will be duplicated as the bacterium divides. The problem thus breaks down into three parts: to find the right structural sequence (insulin's, for example), to place it in bacteria in such a way that it will be maintained as the bacteria grow and then to manipulate the surrounding information, modifying the regulatory commands so that the structural sequence is expressed as protein. Once the protein is made, still further changes in its gene or modifications of the bacterium may be needed to obtain the protein in large enough amounts to be useful.

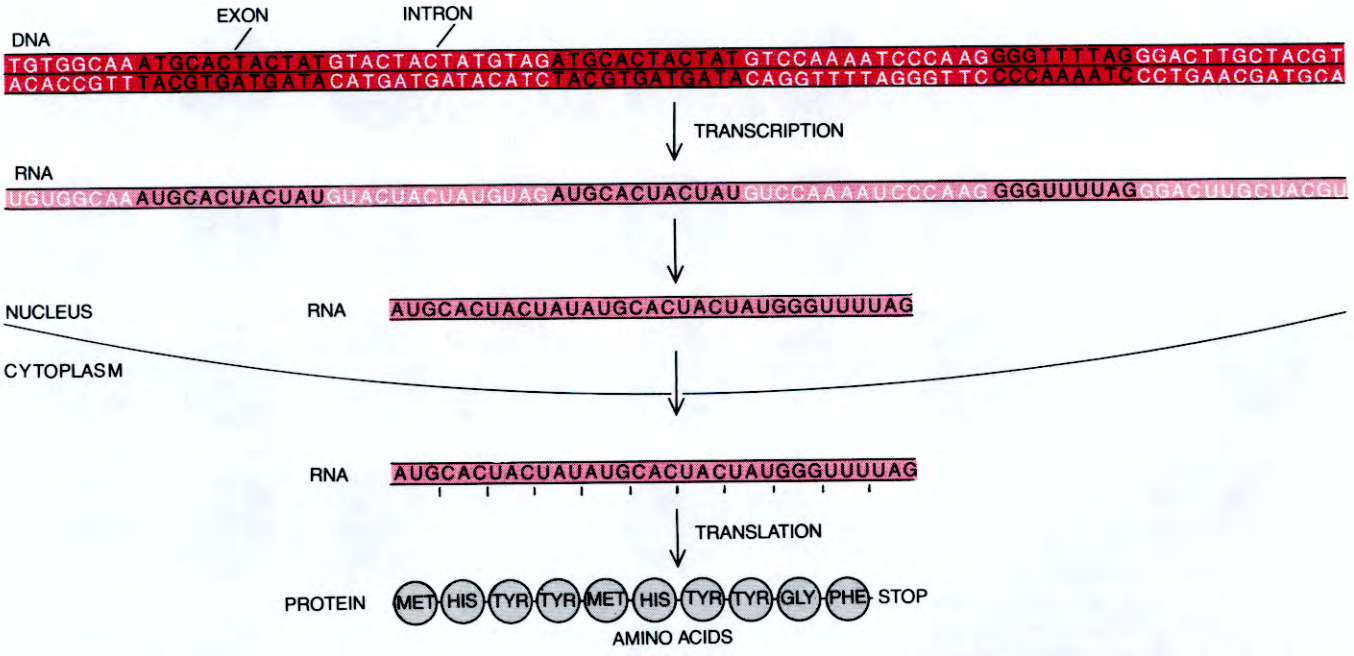
The constellation of recombinant-



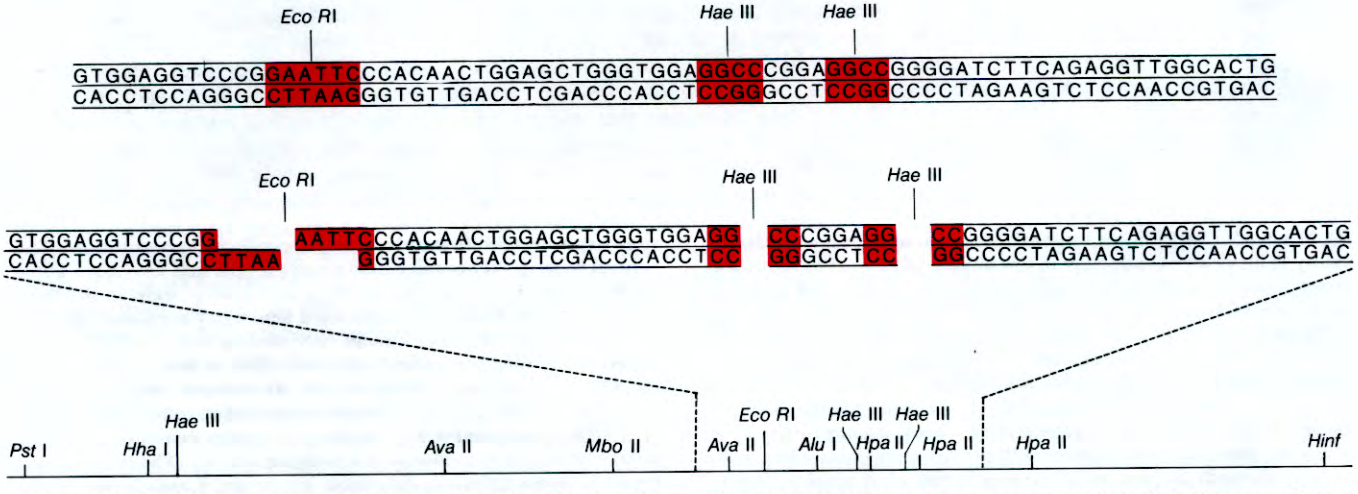
HUMAN INTERFERON synthesized in bacteria demonstrates its ability to block a viral infection in this biological assay. The structural information for making the protein interferon was obtained from human white blood cells in the form of messenger-RNA molecules; the RNA then served as a template for the synthesis of double-strand molecules of copy DNA, and the DNA in turn was inserted by recombinant-DNA techniques into a laboratory strain of the bacterium *Escherichia coli*, which synthesized the protein. For the assay dilutions of an extract of the bacteria were placed in some of the wells of a clear plastic tray; the other wells served as controls. (The wells are seen through the bottom of the tray in this photograph.) Human cells were added to the wells and were grown to form a layer of cells covering the bottom of each well. A virus preparation was then added to the cells. Twenty-four hours later the cell layer was stained. Where interferon in the extracts protected the cells against the virus the cells survived and were stained. Where there was no interferon the virus killed the cells and the dead cells did not pick up the stain. The control wells in the first column at the left contain a layer of cells that were never exposed to the virus; they accordingly appear stained. The control wells in the second column contain cells that have been killed by the virus; they look gray or clear. The control wells in the third column contain dilutions of a standard laboratory sample of interferon obtained directly from human cells; the top well has the most interferon and each succeeding well has a third as much interferon as the well above it. The wells in the next six columns hold dilutions of bacterial extracts from six different colonies of *E. coli* in which interferon DNA was present. Five of the six columns containing the bacterial extracts show evidence of interferon activity. The third extract tested (Column 6) had no detectable interferon; it apparently did not have a complete interferon gene. The synthesis of human interferon by the recombinant-DNA method was achieved by Charles Weissmann and his colleagues at the University of Zurich in collaboration with Kari Cantell of the Finnish Red Cross. The work was supported by Biogen, SA. Interferon is synthesized by many animal cells, but it is species-specific: only human interferon works for human beings, and it has been too scarce even for satisfactory experimentation.

DNA techniques for placing and maintaining a new gene in bacteria is called cloning, which in this sense means the isolation of a specific new DNA sequence in a single organism that proliferates to form a population of identical descendants: a clone. There are two convenient ways of doing this. In one method a small circular piece of DNA called a plasmid is the vehicle for introducing the new DNA into the bacterium. Plasmids carry only a few genes of their own and are maintained in several copies inside the bacterium by the bacterium's own gene functions; they remain separate from the main set of bacterial genes carried on a circle of DNA about 1,000 times larger. Alternatively the vehicle could be a virus that grows in bacteria. Such viruses normally have some 10 to 50 genes of their own (a bacterium has several thousand genes) and can often carry other new DNA segments in place of some of their own. All the techniques we shall describe apply to both plasmids and viruses.

A molecule of DNA resembles a very



PROTEINS ARE MADE in a living cell according to instructions encoded in the cell's genes, which consist of specific sequences of chemical groups (bases) strung out along a double-strand molecule of DNA in the cell's nucleus. The genetic code is "written" in the four letters *A, T, G* and *C,* which stand respectively for the four bases adenine, thymine, guanine and cytosine. The code is "read" in the three-letter sets called codons, which specify the amino acids linked together in the protein chain. The order of the bases can also convey regulatory commands. In multicelled organisms the structural sequence, or gene, encoding a particular protein is usually broken into fragments separated by long stretches of other DNA; in this diagram the gene fragments, called exons, are represented by the black letters and the intervening sequences, known as introns, by the white letters. The genetic information is translated into protein indirectly. First the entire sequence of bases is transcribed inside the nucleus from the DNA to a single-strand molecule of RNA. According to the base-pairing rules governing transcription, adenine always pairs with uracil (*U*) and guanine always pairs with cytosine. Next the RNA copies of the introns are excised from the message and the remaining RNA copies of the exons are joined together end to end. The reassembled strand of messenger RNA then moves from the nucleus to the cytoplasm, where the actual protein-manufacturing process takes place.



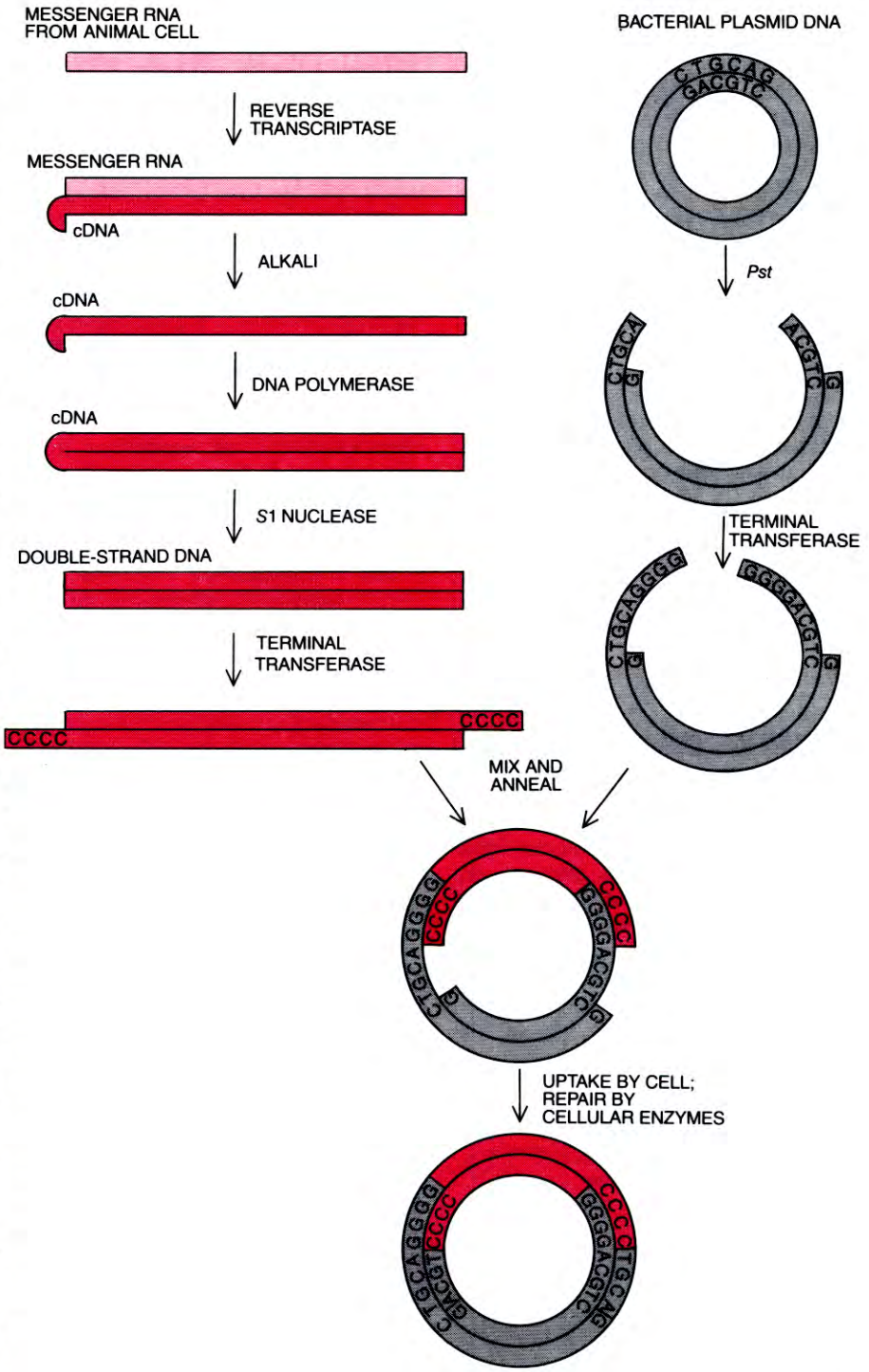
DNA CAN BE CUT into comparatively short lengths with the aid of restriction endonucleases, special enzymes that recognize specific base sequences at which they cause the molecule to come apart. For example, *Eco RI*, the first such enzyme discovered, recognizes a certain six-base sequence and cuts the molecule wherever this sequence appears, whereas *Hae* III, another restriction enzyme, operates at a certain four-base sequence. Since the probability of finding a particular four-base sequence is greater than that of finding a particular six-base sequence, one would expect *Hae* III to cut DNA more often than *Eco RI*. Accordingly one *Eco RI* site and two *Hae* III sites are represented in the DNA segment at the top, which corresponds to part of the gene coding for insulin in rat cells. The same DNA contains recognition sites for a number of other restriction enzymes, as is shown in the line diagram of a larger gene fragment at the bottom.

long, twisted thread. A bacterium has one millimeter of DNA in a continuous string of some three million bases folded back and forth several thousand times into a space less than a micron (a thousandth of a millimeter) across. In human cells the DNA is packed into 46 chromosomes, each one containing about four centimeters in a single piece, the total amount corresponding to about three billion bases. How can one find and work with a single gene only a few thousand bases long? Fortunately nature has devised certain enzymes (proteins that carry out chemical reactions) that solve part of the problem. These special enzymes, called restriction endonucleases, have the ability to scan the long thread of DNA and to recognize particular short sequences as landmarks at which to cut the molecule apart. Some 40 or 50 of these enzymes are known, each of which recognizes different landmarks; each restriction enzyme therefore breaks up any given DNA reproducibly into a characteristic set of short pieces, from a few hundred to a few thousand bases long, which one can isolate by length.

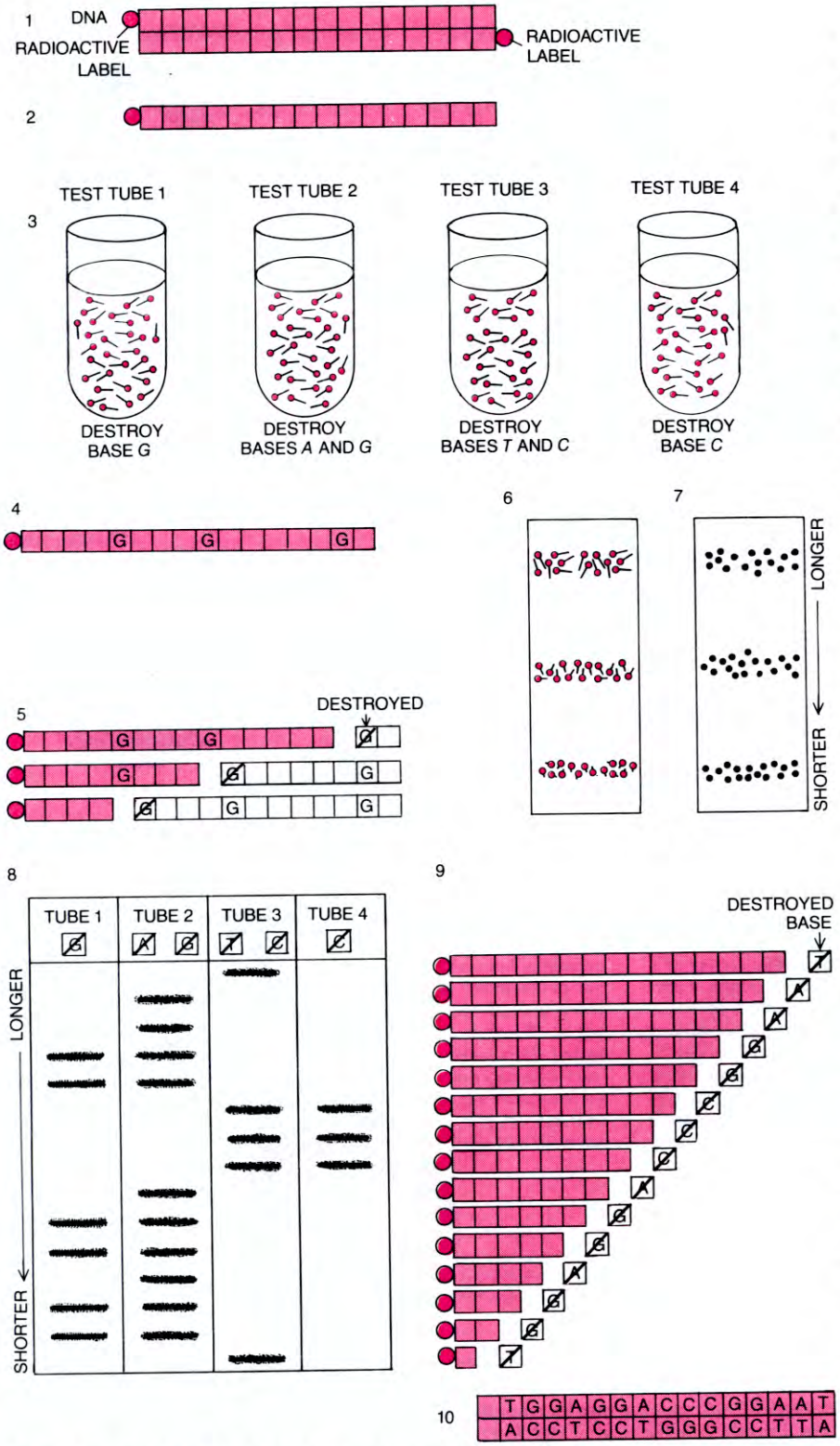
One can clone such DNA pieces in bacteria. As a first step one purifies the circle of plasmid DNA. The sequences of the plasmids are such that one of the restriction enzymes will recognize a unique site on the plasmid and cut the circle open there. One can insert a chosen DNA fragment into the opening by using a variety of enzymatic techniques that connect its ends to those of the circle. Ordinarily this recombinant-DNA molecule could not pass through the bacterial cell wall. A dilute solution of calcium chloride renders the bacteria permeable, however; in a mixture of treated cells and DNA a few bacteria will take up the hybrid plasmid. These cells can be found among all those that did not take up the DNA if a gene on the plasmid provides a property the bacterium must have to survive, such as antibiotic resistance. Then any bacterium carrying the plasmid will be resistant to the antibiotic, whereas all the others will be killed by it. When one spreads the mixture of bacteria out on an agar plate containing nutrients and the antibiotic, each single bacterium with a plasmid will grow into a separate colony of about 100 million cells. A single colony can be chosen and grown further to yield billions of cells, each of which contains identical copies of the new DNA sequence in a recombinant plasmid.

The Sequencing of DNA

The procedures we have outlined so far are followed in "shotgun" cloning experiments. One breaks up the DNA of an animal cell into millions of pieces and inserts each piece into a different bacterium. In this way a number of collections of all the fragments of human,



RECOMBINANT-DNA TECHNIQUE for making a protein in bacteria calls for the insertion of a fragment of animal DNA that encodes the protein into a plasmid, a small circular piece of bacterial DNA, which in turn serves as the vehicle for introducing the DNA into the bacterium. The plasmid DNA is cleaved with the appropriate restriction enzyme and the new DNA sequence is inserted into the opening by means of a variety of enzymatic manipulations that connect the new DNA's ends to those of the broken plasmid circle. In the procedure illustrated here, for example, a special enzyme, reverse transcriptase, is first used to copy the genetic information from a single-strand molecule of messenger RNA into a single strand of copy DNA. The RNA template is then destroyed, and a second strand of DNA is made with another enzyme, DNA polymerase. Still another enzyme, *S1* nuclease, serves to break the covalent linkage between the two DNA strands. In the next step the double-strand DNA is joined to the plasmid by first using the enzyme terminal transferase to extend the ends of the DNA with a short sequence of identical bases (in this case four cytosines) and then annealing the DNA to the plasmid DNA, to which a complementary sequence of bases (four guanines) has been added. Bacterial enzymes eventually fill the gaps in the regenerated circular DNA molecule and seal the connection between the inserted DNA and the plasmid DNA. The particular plasmid used by the authors to make rat proinsulin in bacteria, designated *pBR322*, incorporates two genes that confer resistance to two antibiotics: penicillin and tetracycline. The plasmid is cleaved by the restriction enzyme *Pst* at a recognition site that lies in the midst of the gene encoding penicillinase (the enzyme that breaks down penicillin). The added DNA destroys this enzymatic activity, but the tetracycline resistance remains and is used to identify bacteria containing the plasmid.



SEQUENCING OF DNA, in the method devised by one of the authors (Gilbert) and Allan M. Maxam, begins with the attachment of a radioactive label to one end of each strand of double-strand DNA (*1*). The strands of trillions of molecules are separated (*2*) and a preparation of one of the two kinds of strands is divided among four test tubes (*3*). Each tube contains a chemical agent that selectively destroys one or two of the four bases *A*, *T*, *G* and *C*, thereby cleaving the strand at the site of those bases; the reaction is controlled so that only some of the strands are cleaved at each of the sites where a given base appears, generating a set of fragments of different sizes. A strand containing three *G*'s (*4*), for example, would produce a mixture of three radioactively labeled molecules (*5*). The reactions break DNA at the *G*'s alone, at the *G*'s and the *A*'s, at the *T*'s and the *C*'s, and at the *C*'s alone. The molecules are separated according to size by electrophoresis on a gel; the shorter the molecule, the farther it migrates down the gel (*6*). The radioactive label produces an image of each group of molecules on an X-ray film (*7*). When four films are placed side by side (*8*), the ladderlike array of bands represents all the successively shorter fragments of the original strand of DNA (*9*). Knowing what base or pair of bases was destroyed to produce each of the fragments, one can start at the bottom and read off a left-to-right sequence of bases (*10*), which in turn yields the sequence of the second strand.

mouse, rat and fly DNA have been made. One can determine the structure of any one of these cloned DNA's by breaking up the hybrid plasmid with a restriction enzyme, separating the resulting DNA fragments, determining the base sequence of each of the fragments and then putting the sequences together to deduce the entire structure of the cloned DNA.

There are two methods for sequencing DNA. Both exploit reference points created by restriction-enzyme cleavage of the DNA at a specific short sequence and then work out the rest of the sequence by measuring the distance of each base from that cut. They do this by creating a set of radioactively labeled molecules, each of which extends from the common point to one of the occurrences of a specific base. When these molecules are separated by size and detected by their radioactivity, the length of the smallest one shows the position of the first occurrence of that base; longer molecules correspond to later occurrences. The pattern created by the analysis of these molecules looks like a ladder. From the positions of the rungs one reads off the lengths. By comparing four such patterns one reads off a sequence.

One technique, devised by Allan M. Maxam and one of us (Gilbert), makes use of chemical reagents that detect the different chemical properties of the bases and break the DNA there. To generate the set of fragments the reactions are done for a short time, so that the molecule is broken only occasionally instead of everywhere the base occurs; different molecules will be broken at different places. Four different sets of reagents are used to generate the four patterns. The radioactive label is attached directly to the end of the particular restriction fragment one wants to sequence, so that only the molecules stretching from the labeled end to the break are detected by their radioactivity.

The other sequencing method, devised by Frederick Sanger of the British Medical Research Council Laboratory of Molecular Biology in Cambridge, makes a DNA copy with an enzyme and stops the sequential synthesis, and hence the elongation of the copy, by blocking the movement of the enzyme at a specific base. Here the radioactive label is incorporated into the newly synthesized molecule in four different reactions. Both methods can provide the sequence of from 200 to 300 bases in a single experiment. One of the small plasmids involved in our cloning experiments was sequenced in a year by Gregory Sutcliffe, who worked out the order of the 4,357 bases on one strand and checked them by working out the complementary strand.

Any DNA region carried on a plasmid can be isolated and sequenced. The difficulty is not in determining the sequence but in obtaining the specific

DNA fragments needed. The recombinant-DNA technique serves almost as a microscope to isolate and to magnify, by making many copies, a DNA region, but one does not want to look through a million bacteria to find a specific gene. The fundamental problem, which has no general solution, is to place only the desired DNA sequence—the desired structural gene—in a bacterium.

Getting the Right Gene

One straightforward approach is suitable for very small proteins. The amino acid sequence and the genetic code will predict a sequence of bases that can specify those amino acids. One can then chemically synthesize a corresponding DNA molecule. Exactly this was done by Keiichi Itakura and his co-workers at the City of Hope National Medical Center in Duarte, Calif., who constructed a DNA sequence 42 bases long that dictates the structure of somatostatin, a small hormone consisting of 14 amino acids. The longer the stretch of DNA, however, the harder it is to make; the synthesis of a stretch of DNA 100 bases long is extremely difficult. Many small hormones consist of from 50 to 100 amino acids, and enzymes and other proteins range from 200 to several thousand amino acids in length. Furthermore, one does not know the amino acid sequence of many interesting proteins. (Indeed, the amino acid sequence of some of these proteins has become available only through the sequencing of cloned DNA.)

The desired structural gene is present, of course, somewhere on the DNA of the animal cell. The problem is to find it, but even if that were possible, the structural information would be broken up (as we mentioned above) by long stretches of other DNA. The information does exist in a continuous form, however, on the messenger RNA. Moreover, different cells specialize in the synthesis of different proteins, so that the appropriate tissue will contain the desired messenger RNA along with other messengers for the common proteins made by all cells. Insulin, for example, is made by the beta cells of the pancreas; those cells contain insulin messenger RNA and other cells do not, even though the insulin gene is present in the DNA of every cell.

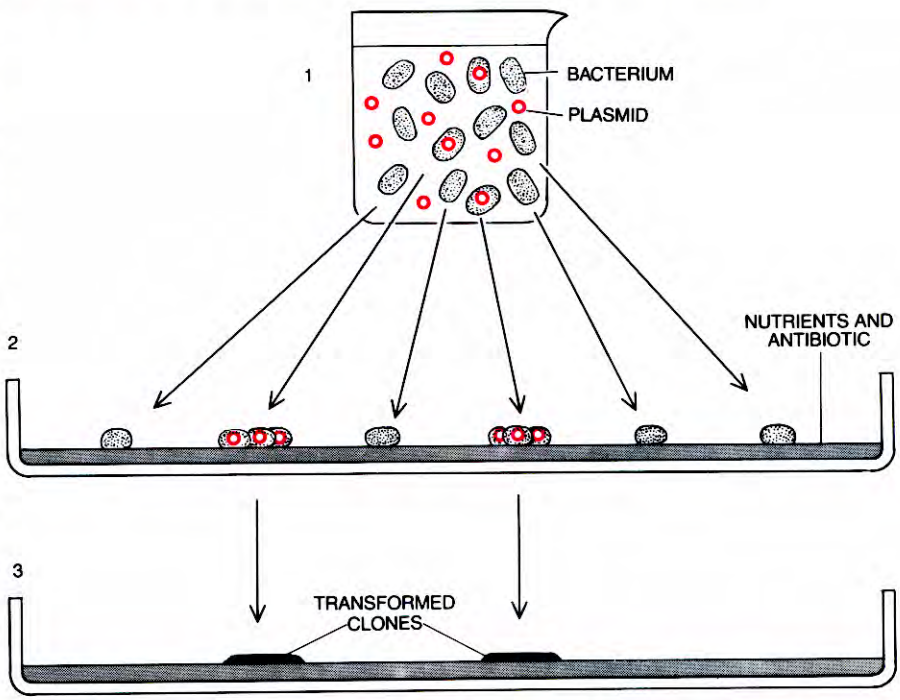
The task is then to convert the desired structural information from the cell's messenger RNA into DNA, which can be cloned. For this one takes advantage of a special enzyme, reverse transcriptase, that can copy a single strand of RNA to make a complementary strand of DNA. (The enzyme is found in certain RNA viruses that reverse the normal DNA-to-RNA transcription. Such viruses depend on RNA rather than DNA to carry their information from one cell to another and convert the RNA back into DNA with the help of reverse transcriptase after they infect a new cell.) One takes this strand of complementary DNA, called copy DNA, and makes a second strand of DNA with the more usual DNA-copying enzyme. The resulting double-strand cDNA fragments are more or less complete copies not only of the desired messenger RNA but also of all the other messenger RNA's that were present in the tissue. At best, however, only a few of the DNA fragments contain all the wanted structural information. Even in those fragments the regulatory signals that surround the structural sequences refer to translation in the animal cell, not in bacteria, and (since the DNA was made from RNA) there will be no transcriptional commands. Although the cDNA can be cloned, two problems remain: to detect any clones containing the sought-after structural DNA fragment and to provide the appropriate signals.

Finding the Right Clone

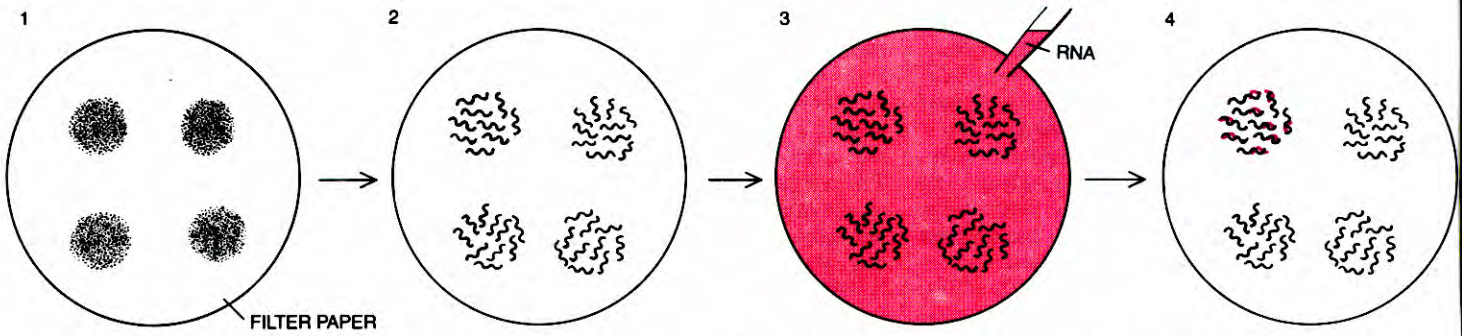
It is simple to find the right clone if the experiment began with a pure messenger RNA. One can detect matching sequences by the process called hybridization. The two strands of a DNA molecule can be separated by heating, which breaks the weak bonds that hold the two strands together without breaking the strong chemical bonds between bases along the chain. When a mixture of such strands is cooled, those sequences that match will find each other. The first step of this process is called denaturation, the second step reannealing. The same process serves to identify sequence matches between RNA and DNA.

One grows bacterial colonies on a disk of cellulose nitrate paper, breaks open the bacterial cells where they lie and fixes the released DNA to the paper. When the DNA is denatured and reannealed to radioactive RNA, only the remains of those colonies that contained a plasmid whose sequence matches the messenger become radioactive. Since one keeps a replica (a living duplicate set of the colonies), one can obtain bacteria containing the desired DNA. One grows these bacteria to provide material to identify, in further hybridization tests, other clones that contain the same sequence in different surroundings and may turn out to be more effective in producing the wanted protein.

If one cannot purify the messenger RNA because the specific messenger is a small fraction of all the messengers in a cell, there are other ways to search for the DNA sequence. One useful property is the detailed shape of the corresponding protein molecule. Those shapes that are most different and distinctive can be recognized by the protein molecules called antibodies. Animals make antibodies as part of their protective response to foreign substances. If one injects human insulin into a guinea pig, for example, the guinea pig will make antibodies that bind to human insulin. These antibodies will not bind to guinea pig insulin because they "see" only the shapes that make the human protein different. A purified antibody, then, can



RECOMBINANT PLASMIDS (*color*) bearing the inserted animal-protein genes and genes for resistance to tetracycline are mixed with bacteria (*1*). Some cells take up the plasmid. The mixture of cells is spread on a culture medium containing the antibiotic (*2*), which kills all the cells that do not have the plasmid. The cells that have taken up the plasmid are antibiotic-resistant; they live, and each of them gives rise to a clone, a colony of genetically identical cells (*3*).



CLONE CONTAINING DESIRED DNA can be found among all the successfully transformed clones (*1*) by means of RNA-DNA hybridization if one has a pure messenger-RNA probe for the desired sequence. The cells are broken open and their DNA is denatured and fixed to filter paper (*2*). The RNA probe (RNA molecules labeled with a radioactive isotope) is added (*3*). The RNA (*color*) will anneal to any DNA whose sequence it matches, forming RNA-DNA hybrids (*4*); the remainder of the RNA is washed away. The presence of the hy-

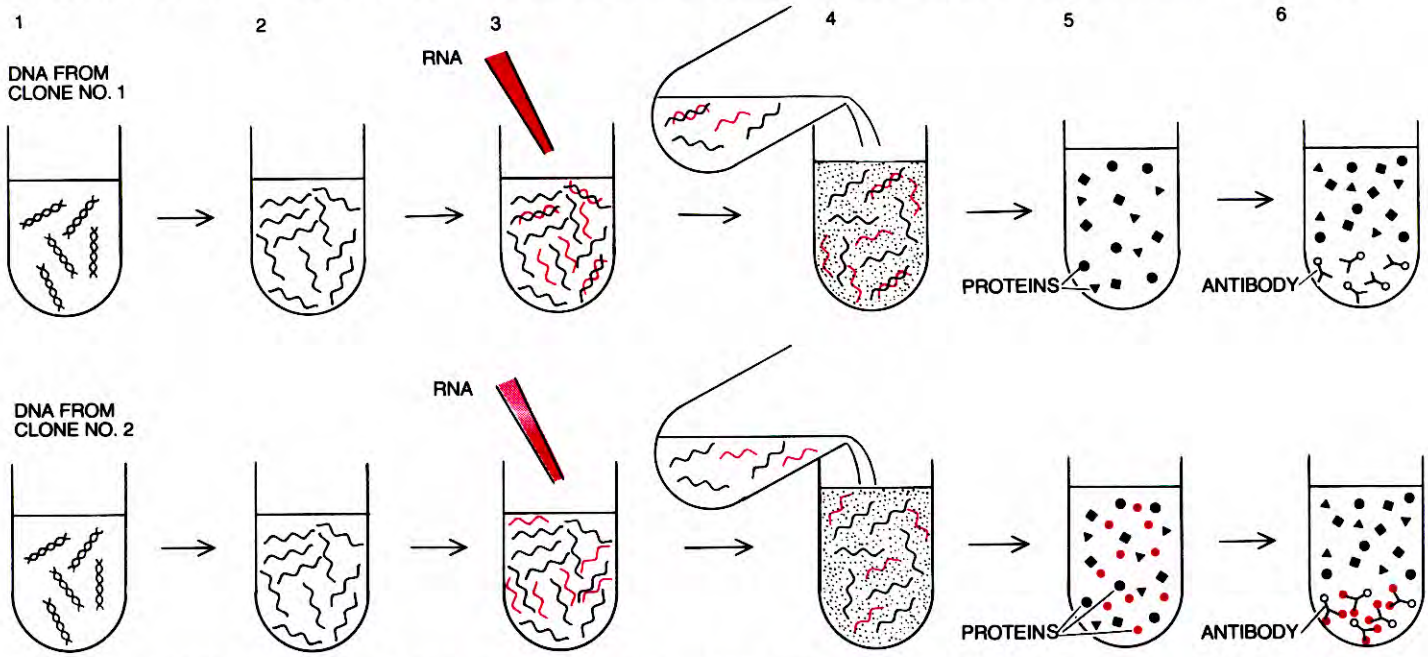
serve as a reagent to detect a particular protein. (This is the way vaccines work. If an animal is injected with an inactivated virus, it is stimulated to make antibodies against the viral proteins. Thereafter the antibodies will protect the animal against infection by that virus by binding to the virus particle and signaling other cells to remove the invader. Without the earlier stimulation the antibody response to the invading virus is too slow to block the infection.)

Even without purifying a specific messenger RNA one can make the RNA molecules function in the test tube by adding the machinery needed to translate the messengers (obtained from the cytoplasm of broken cells) along with radioactive amino acids. Among the small amounts of radioactive proteins that are synthesized one can recognize the protein of interest with antibodies. This provides a means of detecting the presence of a specific messenger. If one takes a recombinant plasmid and hybridizes it to the mixture of RNA's, only the RNA that matches a sequence in the plasmid will anneal to it and therefore no longer function in translation; the plasmid of interest is detected by its ability to block the synthesis of the desired protein. This identification can be verified because the RNA bound to the DNA can be separated from all the other RNA's and then released from the DNA, whereupon it will function to direct the synthesis of the protein.

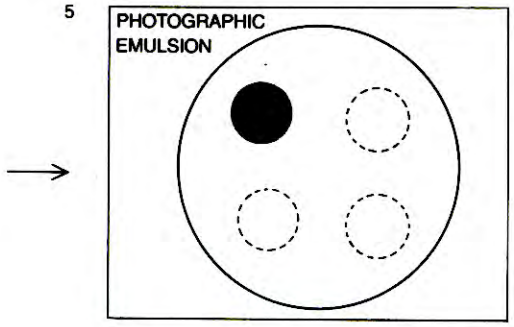
Regulatory Signals

With these techniques one can clone and identify DNA fragments carrying the information that dictates the structure of a protein. Will the information work in bacteria?

One must provide regulatory signals the bacterium can use. One of them is the signal to start the synthesis of a messenger RNA; in bacteria it is a region of DNA immediately in front of the segment of DNA that will be transcribed into RNA. The second important signal functions as part of the messenger RNA, telling the bacterial translation machine to "Start here." All bacterial genes have these two kinds of start signals (some of which work better than others). They also have two stop signals, one for translation and one for transcription. A simple way to make the new protein sequence is to cut a bacterial gene open in its middle with a restriction enzyme and to insert the new DNA there. This results in a hybrid protein that starts out as some bacterial protein and then continues as the string of amino acids one wants. That is how the chemically synthesized gene for somatostatin was made to work in bacteria. The DNA



HYBRID-ARRESTED TRANSLATION, a technique developed by Bryan Roberts of the Harvard Medical School, identifies a clone (*top*) containing the desired DNA even in the absence of a purified RNA probe. DNA from clones being tested (*1*) is denatured (*2*). Unpurified RNA (the same RNA used to make the inserted DNA) is added (*3*); it anneals to any matching DNA. Placed in a "translation system" containing radioactively labeled amino acids (*4*), the unhybridized RNA directs the synthesis of radioactive proteins, but the hybridized RNA cannot be translated; the specific protein (*color*) encoded by the desired DNA is not synthesized in the presence of the clone containing that DNA (*5*). The presence or absence of that protein is determined by an antibody test. Antibody to the protein, fixed to plastic beads,

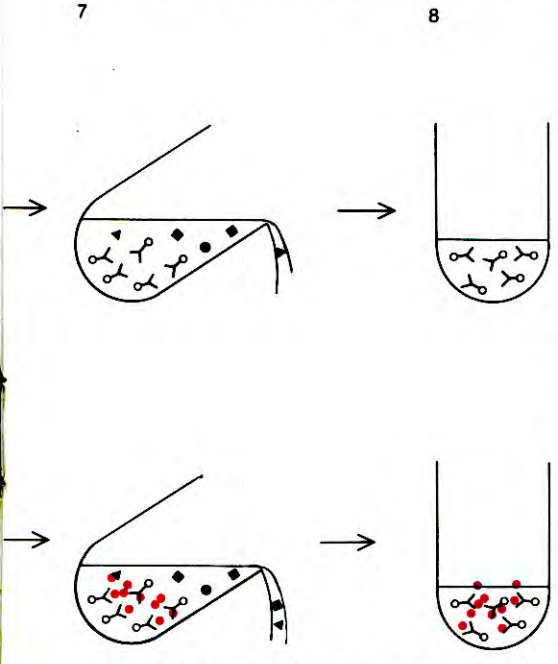


brids is revealed by autoradiography: a photographic emulsion is placed on the filter paper and after exposure the clone containing the desired DNA is identified as a dark spot (*5*).

for those 14 amino acids, followed by a stop signal, was inserted near the end of a 1,000-amino-acid protein. After the bacterium made the hybrid protein the somatostatin part was cleaved off chemically and purified.

Not only can the bacterial gene serve to provide the regulatory signals but also it may endow the hybrid protein with further useful properties. For example, a few bacterial proteins are secreted through the membrane that surrounds the cell. If one inserts the animal DNA into the gene for such a protein, the bacterial part of the hybrid protein will serve as a carrier to move the new protein through the membrane so that it is more easily observed and purified.

We exploited all the techniques described above to obtain a copy of the insulin gene and to insert it into bacteria to make proinsulin. Insulin is a small hormone made up of two short chains, one chain 20 amino acids long and the other 30 amino acids long. These two chains are initially part of a longer chain of 109 amino acids, called preproinsulin. As preproinsulin is synthesized in the beta cells of the pancreas, the first 23 amino acids of the chain serve as a signal to direct the passage of the molecule through a cell membrane. As this happens those amino acids are cleaved off, leaving a chain of 86 amino acids: proinsulin. The proinsulin chain folds up to bring the first and last segments of the chain together, and the central portion is cut out by enzymes to leave insulin. The role of the central portion is to align the two chains comprising insulin correctly. If the two chains are taken apart later, they do not reassemble easily or efficiently. (In spite of these difficulties Itakura and his co-workers synthesized two DNA fragments corresponding to the two chains of human insulin and attached them separately, like somatostatin, to the same large bacterial gene in order to synthesize two separate hybrid proteins in two different bacteria. Then they cut off the two short pieces, purified them and put them together to form insulin.)



is added and binds the protein, precipitating the protein out of the solution (*6*), which is poured off (*7*). Measurement of the precipitates' radioactivity (*8*) shows that one clone (*top*) contains the desired DNA, because it blocked the synthesis of the specific protein.

The Proinsulin Experiment

In our experiments we started with a tumor of the insulin-producing beta cells of the rat. (We worked with rat insulin because at the time we began our experiments the guidelines established by the National Institutes of Health for recombinant-DNA investigations would not allow us to insert the human insulin gene into bacteria; that prohibition has since been removed.)

We made DNA copies of the beta-cell messenger RNA and put them into a plasmid, in the middle of a gene for a bacterial protein, penicillinase, that would be secreted through the membrane of the bacterial cell. We looked among the bacterial colonies by hybridization, we proved that we had the right hybrid plasmid by blocking the synthesis of insulin in a test tube as we described above and we sequenced the DNA to see exactly what part of the insulin gene we had. Once we had found one hybrid plasmid, we used it to find 48 more by repeating the hybridization test. These 48 clones represented 2 percent of all the clones we had made.

Would any of those clones actually synthesize insulin? We looked among the clones containing insulin DNA for any that were synthesizing a hybrid protein part of which was proinsulin. For this we relied on a sensitive radioactive-antibody test. We coated plastic disks with antibody directed against either insulin or penicillinase and exposed them to the contents of cells from each clone. Any insulin (or penicillinase) present in the cells binds to the antibody and is thereby fixed to the plastic disks. Then we applied radioactively labeled anti-insulin antibody to detect the presence of proteins with insulin shapes. One clone gave positive responses, both on disks coated with anti-insulin and on those coated with antipenicillinase, to radioactive antibody to insulin, thereby demonstrating the presence of a penicillinase-insulin hybrid protein.

To see if the bacteria were secreting the hybrid protein we grew the clone in liquid culture and tried to extract the protein by a method that does not burst the bacterial cell membrane. The test showed the fused protein to be present outside the membrane: it was secreted, as we had hoped it would be.

Sequencing the DNA showed that the DNA fragment and the details of the fusion were such that the structural information in the clone was only for proinsulin and did not contain the "pre" region. In order to make insulin we removed most of the bacterial protein and the middle segment of the proinsulin with the digestive enzyme trypsin. Would the insulin made from the bacteria be an active hormone? Stephen P. Naber and William L. Chick of the Elliot P. Joslin Research Laboratory in Boston tested the molecule by showing that it affected the metabolism of sugar by fat cells, as it should.

Improving the Yield

The amount of proinsulin made by the original clone was very small; we are currently engaged in various manipulations to improve the yield. Regulatory signals must be not only efficient but also optimally placed. One need not be satisfied with the signals that happen to surround preexisting bacterial genes. With restriction enzymes one can clip out small DNA fragments that carry only the regulatory signals and tie them together with a DNA-linking enzyme to make new combinations. One can trim back the ends of these fragments by nibbling off bases with still other enzymes before reconnecting them. This will alter the spacings between the signals and the structural sequence. Although each of these manipulations generates only a small number of correct molecules, by cloning after each step one can make large amounts of the DNA and work out its sequence, and then continue the tinkering.

Moreover, one can synthesize short desired DNA sequences and tie them to other fragments. For example, David V. Goeddel and his co-workers at Genentech, Inc., took a piece of DNA containing the structural information for human growth hormone (168 amino acids), connected it to a synthetic piece of DNA containing part of the translational start signal and attached that combination in turn to a fragment containing the rest of the regulatory signals. When this DNA construction was cloned, the

bacteria made a protein of the shape (as recognized by antibodies) and size of growth hormone (although not yet with demonstrated hormone activity).

Although we do not yet know the optimal combinations of the DNA elements for making insulin in bacteria, finding them is only a matter of time. There are other problems to be considered. Often the new animal proteins are broken down in the bacterial cell because their structure is such that enzymes normally present in the bacteria can digest them. Ways have to be found to stabilize the proteins either by removing these enzymes, by embedding the new protein in a hybrid protein to protect it or by secreting it from the cell. Messenger-RNA molecules themselves are often unstable within the cell; modifications in their structure and in the cell itself can make them more effective and lead to increased protein synthesis. And if the number of copies of the plasmid carrying the gene in each cell can be increased, more of the product will be made.

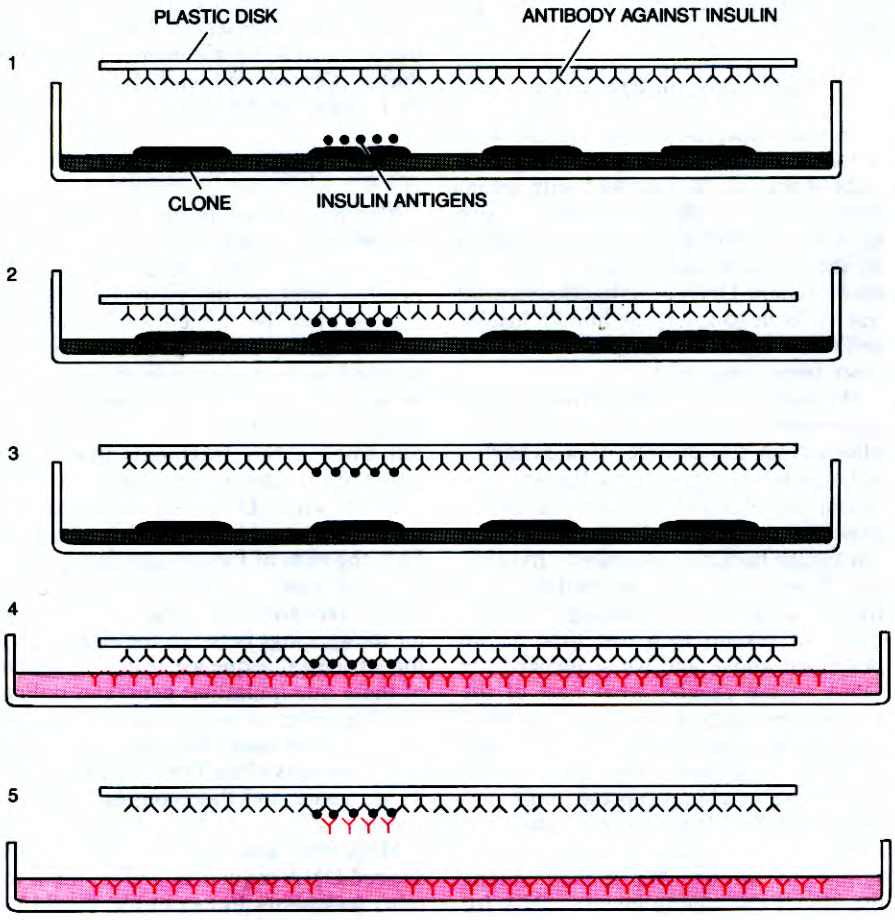
While we work to improve the yield of rat proinsulin and to purify it we expect to apply the same methods to the bacterial synthesis of human insulin. Investigators in other laboratories are also working on the problem, and one can hope that eventually the manufacture of human insulin by bacteria will be cheaper than the purification of insulin from pigs and cattle, the present sources of the hormone. Clearly other human hormones can also be prepared by these procedures. What other therapeutic proteins might be made in bacteria? In general any human protein that cannot be obtained in useful form from animals is an excellent prospect.

Other Proteins from Bacteria

Many genetic diseases are caused by the lack of a single protein. Replacement therapy may be possible if such proteins can be made in bacteria. Vaccines against viral or parasitic infections are a further wide class of possibilities. Today in order to make a vaccine one must be able to grow the disease organism in large amounts; often this is impossible or dangerous. Furthermore, the vaccine must be rendered harmless before it is administered, which can be difficult. The new technology offers the chance to make in bacteria only the protein against which the antibody response needs to be directed. This would eliminate any need to work with the intact disease organism. For example, the hepatitis *B* virus, which causes serum hepatitis, cannot be grown outside the body. The only source of this small DNA virus is the blood of infected human beings. The DNA of the virus has now been cloned in several laboratories and its complete sequence has been worked out, revealing the structure of the viral proteins; now the proteins are being made in bacteria. A flood of new information has resulted from this work.

A particularly promising candidate is interferon, a protein cells make to block viral infections quickly. (The antibody response is much slower.) Interferon appears to be the body's first line of defense against viruses. It may also have a therapeutic effect in some cancers. Interferon has never been available in sufficiently large amounts, however, to determine how effective it might really be in protecting against disease. The ability to test the activities of human interferon will soon be a reality because the protein has now been made in bacteria. Weissmann, with his colleagues Shigekazu Nagata, Hideharu Taira, Alan Hall, Lorraine Johnsrud, Michel Streuli, Josef Ecsödi and Werner Boll, along with Kari Cantell of the Finnish Red Cross, applied many of the techniques we have described to clone and to express this protein. The problem they faced was that the messenger RNA for interferon is far rarer than the one for insulin, even in white blood cells that have been stimulated by infection with a virus to make interferon. They took messenger RNA from these white blood cells (17 liters at a time), made double-strand cDNA and cloned it by the procedures we have described.

They looked through some 20,000 clones (in batches) by hybridizing the plasmid DNA from the clones to the messenger RNA of the white blood cells, isolating the RNA that annealed and checking the RNA to see if it was able to direct the synthesis of interferon (not in the test tube but by injecting the RNA into a particularly large cell, a frog's egg). Fortunately interferon is a remarkably potent substance, and so the amount synthesized in the frog's egg could be detected by its ability to protect cells against viruses.



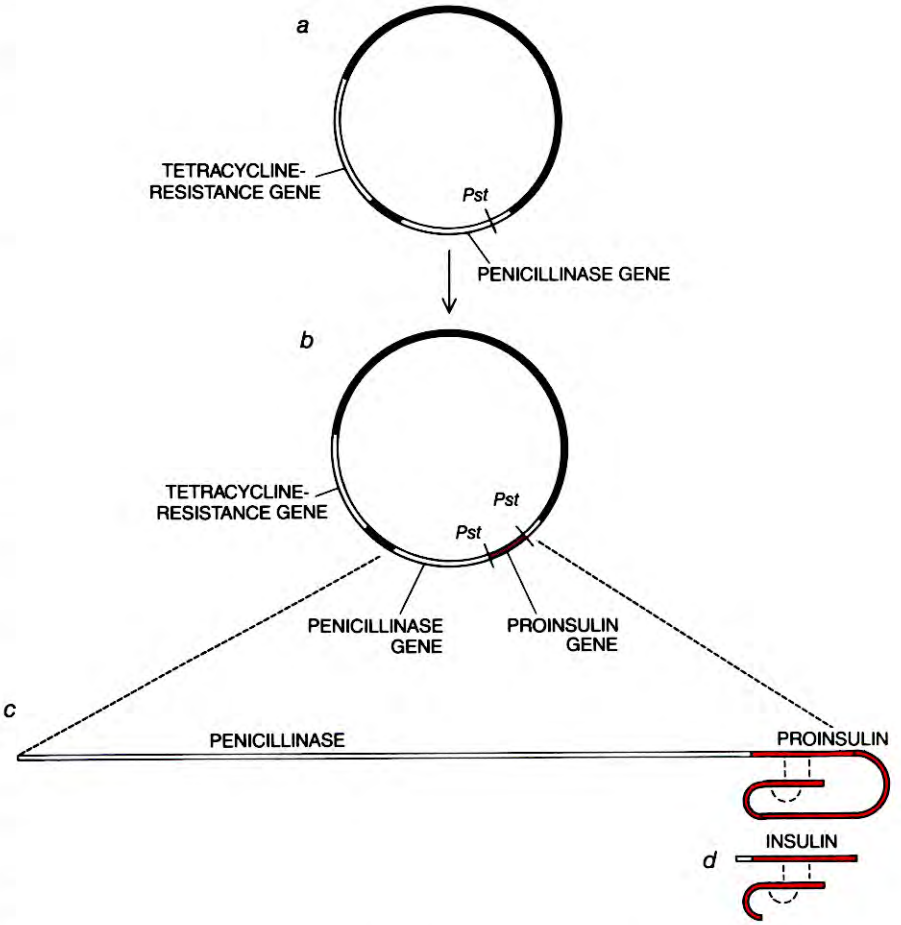
RADIOACTIVE-ANTIBODY TEST, developed by Stephanie Broome and one of the authors (Gilbert), is used to search among the bacterial clones containing insulin DNA for signs that insulin is indeed being synthesized. A plastic disk coated with an anti-insulin antibody is first exposed to the contents of cells from each clone (*1*). Any insulin present in the cells is bound to the antibody (*2*) and thereby fixed to the plastic disk (*3*). Radioactively labeled antibody (*color*) to insulin is then applied to the disk in order to detect the presence of the protein (*4, 5*). When the test is repeated with a plastic disk coated with an antipenicillinase antibody, only a hybrid protein, part penicillinase and part insulin, will bind the labeled antibody.

Once Weissmann and his colleagues had found a batch of clones that could hybridize to interferon messenger RNA they tested progressively smaller groups of those clones to find the correct one. Then, with that clone as a probe, they found other clones by means of hybridization testing. Finally they tested extracts of the bacteria carrying the interferon DNA (inserted into the penicillinase gene) directly to see if any of the bacterial clones made biologically active interferon. A number of clones did, confirming that the interferon structural DNA had been correctly identified. The sequencing of the DNA of those clones will determine the structure of interferon, which is still not known.

The amount of interferon made in the bacteria was extremely small: only one or two molecules per cell. (Bacterial proteins are usually made in from 1,000 to 100,000 copies per cell.) We are confident that the methods we have described will solve this problem and lead to the production of enough interferon for clinical tests.

The Recombinant-DNA Debate

The development of the genetic-engineering techniques described in this article was greeted, over the past decade, with both excitement and alarm. The possible benefits of the techniques were obvious, but some people felt there was reason for concern. Biologists called for an evaluation of the possible hazards of this research; the result was an unprecedented national and international effort in which the public, governments and the scientific community joined to monitor research activities. New knowledge about the properties of genes and the behavior of the bacteria used in this work (usually *Escherichia coli*) has led to a steady lessening of these concerns and to a relaxation of the guidelines that once restricted such experiments. In retrospect, with the advantage of hindsight, the concerns about hypothetical hazards seem to have been unwarranted.



RAT INSULIN WAS OBTAINED by the authors from a hybrid protein composed of part of the bacterial penicillinase molecule and a molecule of proinsulin, an insulin precursor. The map of the plasmid that served as a vehicle, *pBR322* (*a*), shows the location of the genes for the two enzymes conferring antibiotic resistance and the site of cleavage by the restriction enzyme *Pst*. The next map (*b*) shows the structure, as determined by DNA sequencing, of the recombinant plasmid in the bacterial clone that synthesized proinsulin. The proinsulin sequence (*color*) lies between two *Pst* sites that were regenerated in the insertion process. The hybrid protein synthesized by the clone (*c*) comprises most of the penicillinase and also the proinsulin molecule (*color*); broken lines represent disulfide bonds. The authors cut away most of the penicillinase and the middle segment of the proinsulin (*light color*) to make biologically active insulin (*d*).

We know of no adverse effects from this research. The great potential of the new techniques, both in promoting the growth of basic knowledge and in making possible the synthesis of products of direct benefit to society, is much closer to realization than seemed likely only a few years ago.

The Authors

WALTER GILBERT and LYDIA VILLA-KOMAROFF have collaborated on the development of techniques for the enzymatic manipulation of DNA molecules. Gilbert is American Cancer Society Professor of Molecular Biology at Harvard University. A Harvard graduate, he obtained his D.Phil. in mathematics from the University of Cambridge in 1957. He began his career as a theoretical physicist but switched to experimental work in molecular ge-netics about two decades ago. Gilbert is a founder of Biogen, SA, an applied-genetics company. Villa-Komaroff is assistant professor of microbiology at the University of Massachusetts Medical Center. She was graduated from Goucher College in 1970 and received her Ph.D. in cell biology from Harvard in 1975.

Bibliography

A BACTERIAL CLONE SYNTHESIZING PROINSULIN. Lydia Villa-Komaroff, Argiris Efstradiadis, Stephanie Broome, Peter Lomedico, Richard Tizard, Stephen P. Naber, William L. Chick and Walter Gilbert in *Proceedings of the Academy of Sciences of the United States of America,* Vol. 75, No. 8, pages 3727–3731; August, 1978.

SYNTHESIS IN *E. COLI* OF A POLYPEPTIDE WITH HUMAN LEUKOCYTE INTERFERON ACTIVITY. Shigekazu Nagata, Hideharu Taira, Alan Hall, Lorraine Johnsrud, Michel Streuli, Josef Escödi, Werner Boll, Kari Cantell and Charles Weissmann in *Nature,* in press.

THE SYNTHESIS OF EUKARYOTIC PROTEINS IN PROKARYOTIC CELLS. Lydia Villa-Komaroff in *Gene Structure and Expression,* edited by D. H. Dean, L. F. Johnson, P. C. Kimball and P. S. Perlman. Ohio State Press, in press.