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.

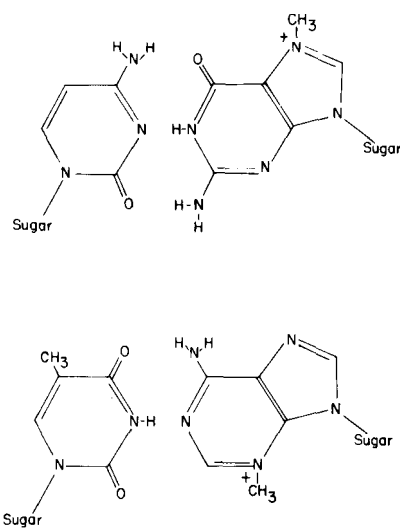


Figure 1. Methylated cytosine-guanine and thymine-adenine base pairs. The top of the figure shows cytosine-guanine base pair methylated at the N7 position of guanine. The bottom of the figure shows a thymine-adenine base pair methylated at the N3 position of adenine. The region above each of the base 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

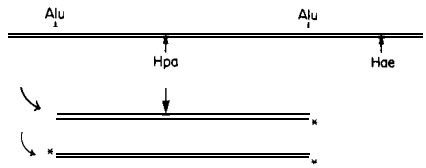


Figure 2. Procedure for obtaining a double-stranded DNA fragment uniquely labeled at one end of one strand. The figure shows the restriction cuts for the enzymes AluI (target AG/CT) and Hpa (target C/CGG producing an uneven end) in the neighborhood of the *lac* operator. The *lac* repressor is shown 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.

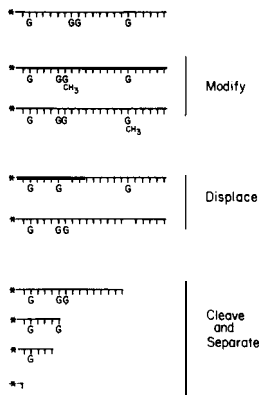


Figure 3. Outline of procedure to produce fragments of DNA by breaking the DNA at guanines. Consider an end-labeled strand of DNA. We modify an occasional guanine by methylation with dimethylsulfate. Heating the DNA will then displace that guanine from the DNA strand, leaving behind the bare sugar; cleaving 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-

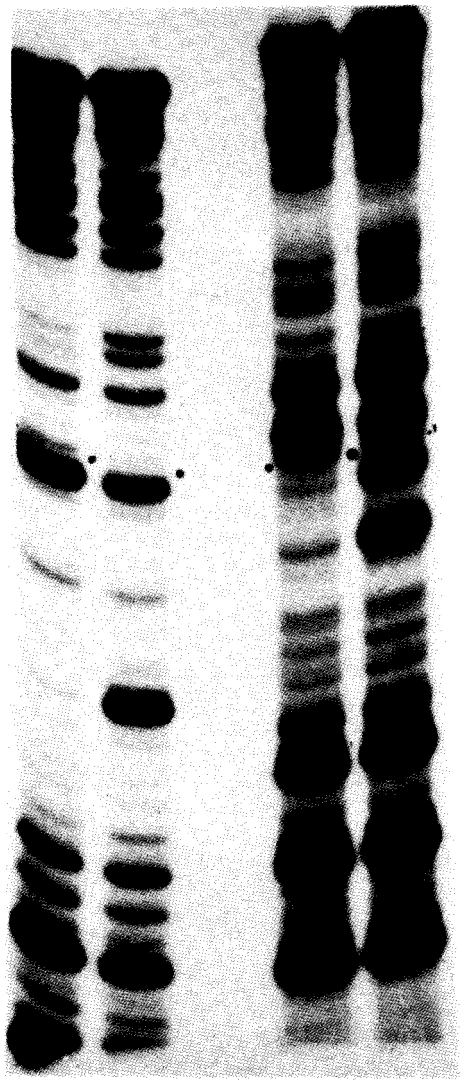
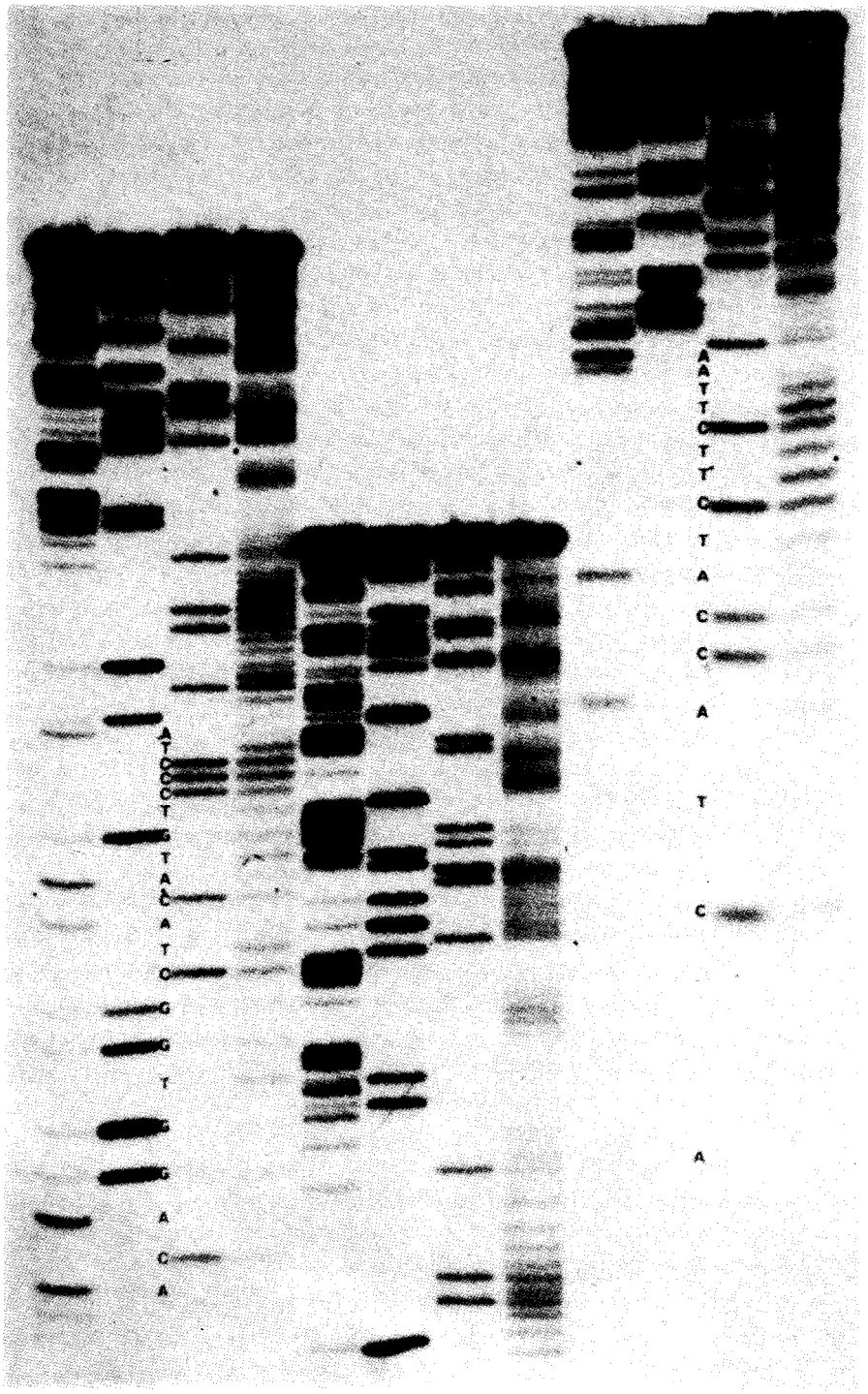


Figure 4. Methylation protection experiment with the *lac* repressor. The columns show the pattern of clevage along each strand of the 53-55 base long fragment bearing the *lac* operator. The second column from the left represents the DNA (labeled at the 5′ end of the 53 base long strand) treated by 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
T
C
C
T
G
T
A
C
A
T
C
G
G
T
G
G
A
C
A
A
A
T
T
C
T
T
C
T
A
C
C
A
T
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).

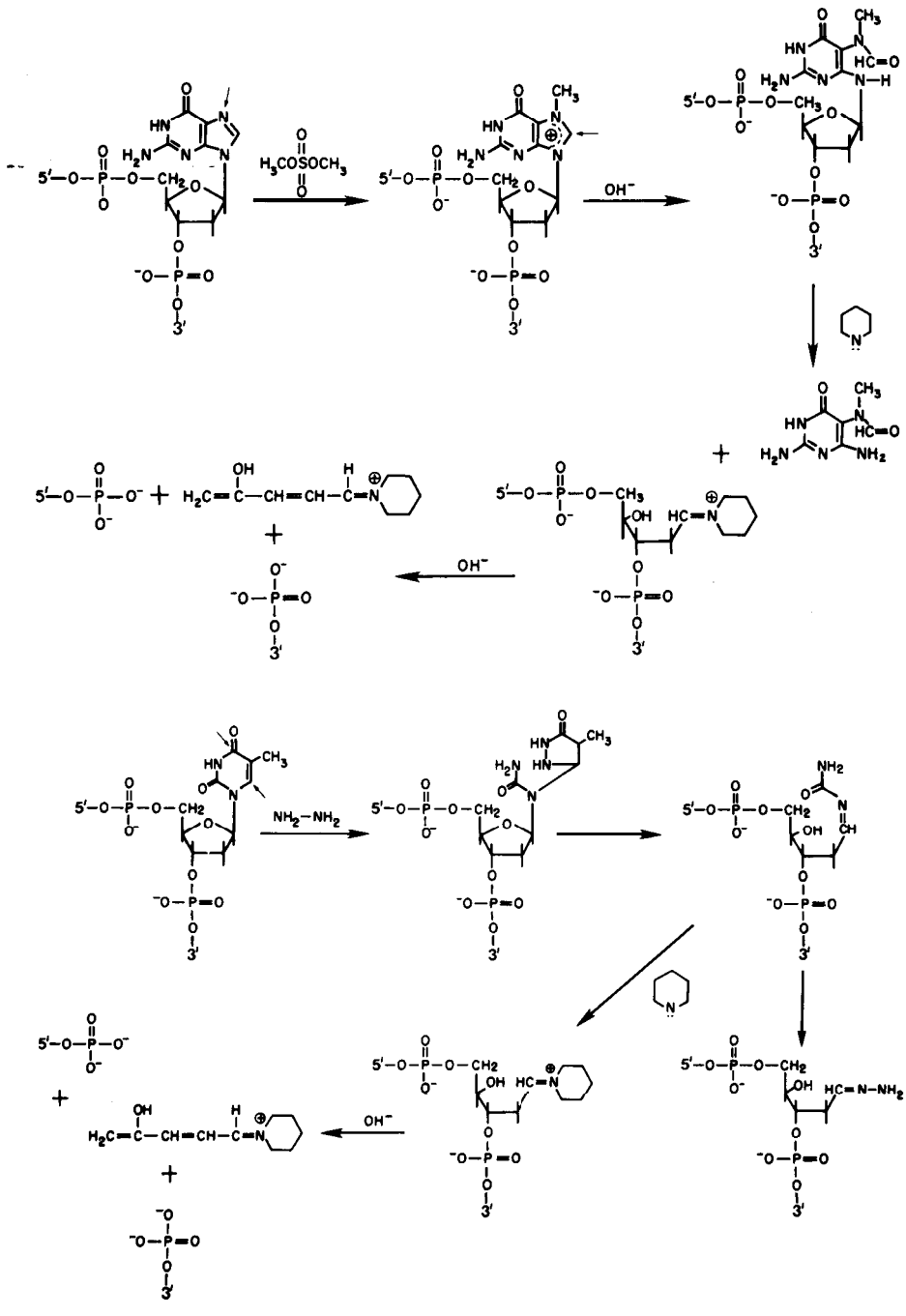


Figure 6. Examples of the detailed chemistry involved in breaking the DNA. Figure 6a. above, shows the guanine breakage. The guanines are first methylated with dimethylsulfate. The imidazole ring is opened by treatment with alkali (during the piperidine treatment). Piperidine displaces the base and then 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.

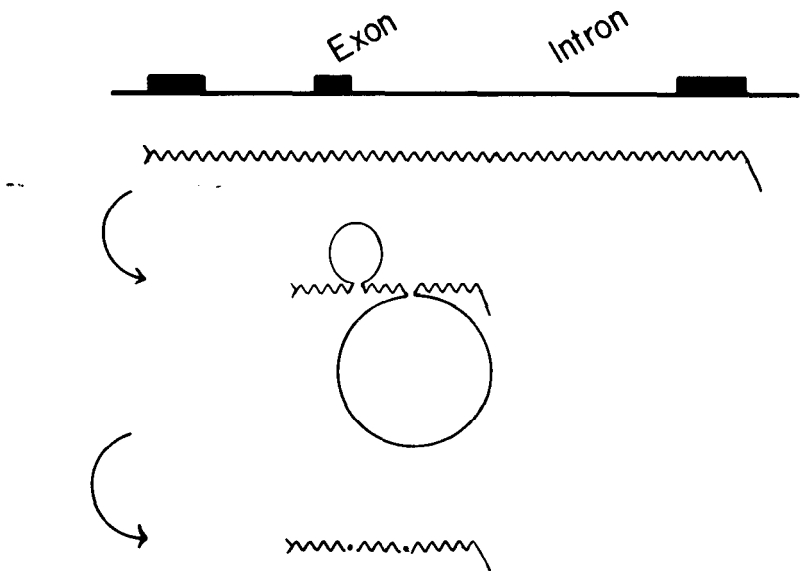


Figure 7. A transcription unit corresponding to alternating exons and introns. The whole gene, a transcription unit, is copied into RNA terminating in a poly (A) tail. The regions corresponding to introns are spliced out leaving a messenger RNA made up of the three exons, the regions that are expressed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

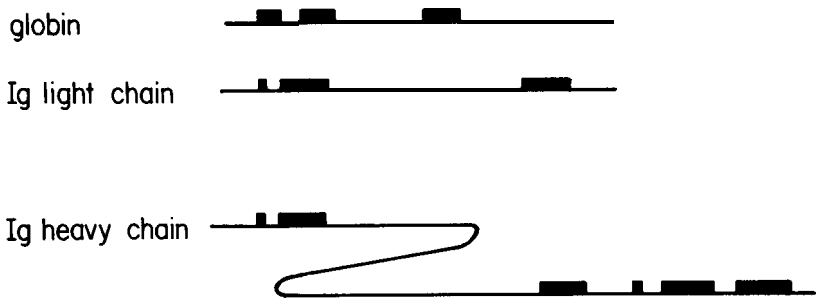


Figure 8. Examples of the intron-exon structure of a few genes. (1) The gene for globin is broken up by two introns into three exons (20). (2) The functional gene in a myeloma cell for the immunoglobulin lambda light chain is broken up into a short exon corresponding to the hydrophobic leader sequence, 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

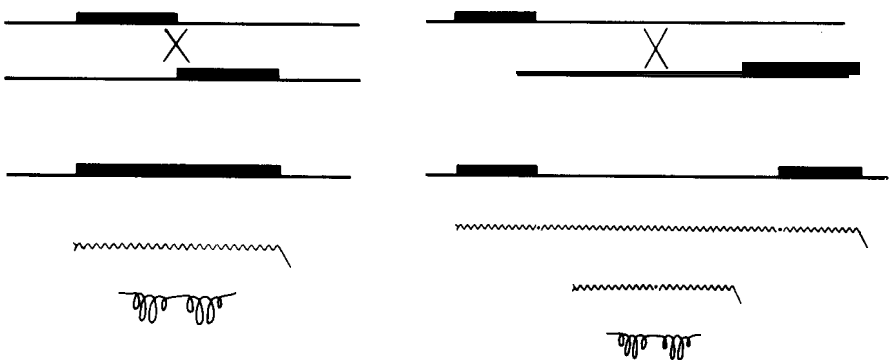


Figure 9. A double-length gene product arises through unequal crossing over. On the left, figure 9a, is the classical process by which a gene corresponding to a single polypeptide chain might have its length doubled by a crossing over. The top two lines indicate two coding regions, brought into accidental 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

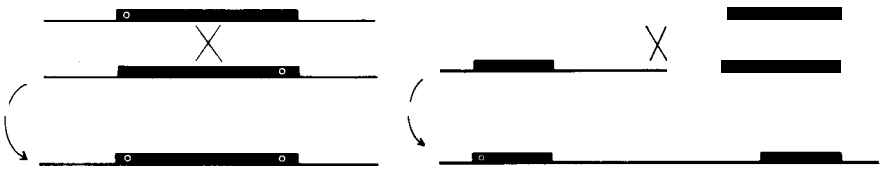


Figure 10. Introns speed legitimate recombination. Figure 10a, on the left, shows the classical pattern by which two mutations, one occurring in one copy of the gene, at the left end, and the other occurring in the other copy of the gene, at the right hand end, might get together by recombination happening 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 "minig-lobin"; the side exons have provided polypeptide material to stablize the protein.

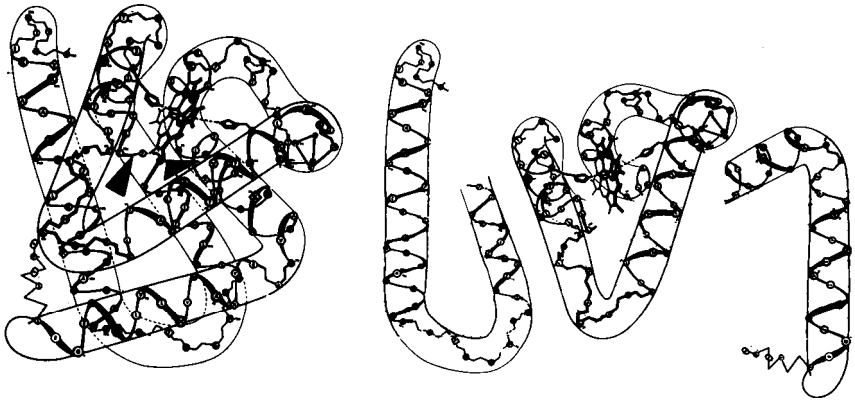


Figure 11, A schematic dissection of globin into the product of the separate exons. At the left the black arrows show the points at which the structure of a chain of globin is interrupted by the introns. (The structures of the chains of globin and of myoglobin are very similar, the schematic structure 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 Opers" *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).