Split Genes and RNA Splicing

Francis Crick

Summary. A number of genes in higher organisms and in their viruses appear to be split. That is, they have "nonsense" stretches of DNA interspersed within the sense DNA. The cell produces a full RNA transcript of this DNA, nonsense and all, and then appears to splice out the nonsense sequences before sending the RNA to the cytoplasm. In this article what is known about these intervening sequences and about the processing of the RNA is outlined. Also discussed is their possible use and how they might have arisen in evolution.

In the last 2 years there has been a mini-revolution in molecular genetics. When I came to California, in September 1976, I had no idea that a typical gene (*1*) might be split into several pieces and I doubt if anybody else had. By the time of the annual Cold Spring Harbor Symposium, in the summer of 1977, it was clear that there was something very strange about the arrangement of the genes in several mammalian viruses, and for this reason it seemed highly likely that some chromosomal genes would also be in several pieces. This has since been found to be the case. Even though the experimental evidence is still very patchy, it is now universally accepted that a gene in a higher organism, coding for a protein, may have other base sequences interspersed within it.

This article is not a historical account of the discovery. The earliest publications can be tracked down by consulting the Cold Spring Harbor volume on chromatin (*2*). Nor does it attempt to be a comprehensive review, both because that would take up too much space and also because experimental results are coming in all the time. I present here an overall view of the present position, fluid though it is, together with some general ideas and a few remarks about

The Basic Problem

It is easiest to begin by considering an imaginary example. The upper part of Fig. 1 shows schematically a length of DNA which codes for a single messenger RNA (mRNA). The lower part shows the base sequences found in that mRNA. The relation between the two is indicated by the lines connecting them. The figure shows that in this case there are two fairly long stretches of base sequence along the DNA of the gene which do not appear in the final mRNA. Such sequences are now known as intervening sequences. An alternative terminology, used by Gilbert and his colleagues (*3*, *4*) refers to the intervening sequences as "introns"; those base sequences on the DNA which do end up in the mRNA are referred to as "exons" since they are the ones which are expressed. At this stage, any terminology is likely to lead, before long, to difficulties and complications (*5*). In this article I use the intron-exon terminology, if only for want of a better one.

What possible mechanisms are there which could have produced this result? There are at least four that immediately spring to mind:

The author is Kieckhefer Research Professor at the Salk Institute, San Diego, California 92112.

0036-8075/79/0420-0264$02.00/0

1) The DNA in the cell producing the message might be rearranged to displace or eliminate the sequences which are not needed. On this hypothesis the DNA in the germ line would remain unaltered.

2) The DNA would remain unaltered but the RNA polymerase, producing the primary RNA transcript, would skip across the introns on the DNA so that only the exons appear in the primary transcript.

3) Each exon would be transcribed separately, and the separate pieces of RNA would then be joined together in the correct order to form the final mRNA.

4) The RNA polymerase would make a primary transcript of the whole region, both exons and introns. This transcript would then be processed so that the introns were removed while at the same time the exons were all joined together in the correct order. This mechanism, which is almost certainly the one that occurs in the majority of cases, is now popularly known as "splicing."

What does the experimental evidence suggest? It has been shown that the first mechanism—the rearrangement of DNA—does indeed occur in one system. A light chain of the immunoglobulin of the mouse (either κ or λ) is coded, in the germ line, on two widely separated stretches of DNA. These are found to be much closer together in the DNA of the somatic cells producing the protein (*2*, *6*, *7*). This is a very important result but I shall not pursue it further. There are good reasons for suspecting that the immune system may be a special case, although not necessarily a unique one. So far there is no evidence at all suggesting that either the second or the third mechanism listed above is actually used (*8*). On the other hand, the evidence (not described in detail here) for the fourth mechanism, is now so widespread that there is little doubt that it, or something very like it, is actually happening. For the rest of this article, therefore, I shall ignore the first three mechanisms and concentrate on splicing.

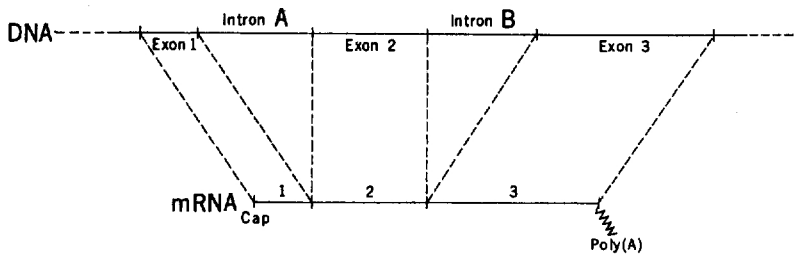

Fig. 1. The top horizontal line represents a stretch of DNA in the genome, the bottom one the mRNA produced from it. In this imaginary example the gene has three exons, marked 1, 2, and 3, and two introns (intervening sequences) lettered A and B. There are no sequences in the mRNA corresponding to those in the two introns.

How Widespread Is Splicing?

I have spoken as if splicing only occurred in the processing of mRNA, but we already know that at least two other species of RNA are spliced. Indeed, one of the earliest discoveries was that some of the transfer RNA (tRNA) molecules in yeast are spliced, although their introns are fairly small (*9*, *10*). More recently two groups of investigators have isolated a crude enzyme preparation that will perform that operation in the test tube (*11*, *12*). The single gene for ribosomal RNA (rRNA) in a yeast mitochondrion appears to contain an intron (*13*). Some genes for rRNA in *Drosophila* also appear to contain introns (*14*), but recent evidence suggests that these particular genes may not be transcribed (*15*). Whether the nuclear precursor of rRNA is ever spliced remains to be discovered. So far, there is no evidence at all to show whether or not other kinds of RNA molecules, such as the small RNA's found in the nucleus, are produced by splicing. Thus splicing is defined as the mechanism by which a single functional RNA molecule is produced by the removal of one or more *internal* stretches of RNA during the processing of the primary transcript.

Where are split genes found? So far, they have only been noticed in eukaryotes. If they were common in prokaryotes (the bacteria and the blue-green algae), they would almost certainly have been discovered earlier. We cannot yet say categorically that they do not occur in prokaryotes but it certainly seems unlikely that they do. They are common in eukaryotic viruses. Indeed, that is where their importance was first realized, but an interesting distinction exists. They have only been found in DNA viruses that occur in the cell nucleus (*2*) or in RNA retroviruses which have a DNA nuclear phase (*16*). Split genes have not so far been discovered in viruses that exist only in the cytoplasm of a cell.

All this would suggest that the phenomenon of splicing is correlated with the existence of a nuclear membrane. This hypothesis would make very good sense. In a prokaryotic cell, which lacks a nuclear membrane, the translation of the message by ribosomes starts well before the transcription of the message from the DNA has finished. In a eukaryotic cell, by contrast, the process of transcription takes place in the nucleus, whereas the process of translation on the ribosomes takes place mainly if not entirely in the cytoplasm. The two operations are separated by the nuclear membrane, and this gives an obvious opportunity for additional processing to take place. Such a hypothesis would predict that split genes would not be found in mitochondria. Unfortunately, the experimental evidence suggests that there, too, genes are split into pieces. In yeast mitochondria the single gene for the larger rRNA molecule is almost certainly split (*13*). The evidence that an mRNA is also split is not yet completely decisive, but it is certainly very suggestive (*17*). For two recent reviews see (*18*). Of course, the enzymes required for splicing will be available in the cell, so it would not be too surprising if they (or a close relative of them) penetrated into the mitochondrion. What is surprising is that there appears to be no evidence that there is any membrane separating the DNA of the mitochondrion from its ribosomes, although yeast may be a special case. This problem is discussed again below.

Splicing in Higher Organisms

I now attempt to give a rapid and necessarily incomplete summary of the distribution of split genes in higher organisms. In mammals, one or more of the globin genes have been shown to be split in several species (*19–25*), as have the genes for certain κ and λ light chains of immunoglobulin in the mouse (*6*, *7*, *26–28*) and the heavy chain of a mouse immunoglobulin (*29*). As was mentioned above, split genes are common in a number of mammalian viruses (*2*). The ovalbumin gene in chickens has been shown to be split into many pieces (*30–38*), and there is suggestive evidence that this is also true for the chick ovomucoid gene (*39*). So far, there is no evidence from other vertebrates, nor for any gene in a plant. There is a report (*40*) that the gene for silk fibroin in the silkworm is split but

there is no definitive evidence for a split gene in *Drosophila* since the ribosomal genes mentioned above may not be transcribed (*15*). If split genes do exist in *Drosophila*, one would expect that a case would be discovered fairly soon. In the fungi, the only example known is that of several tRNA's in yeast (*9–12*). It is obviously impossible to deduce much from such sparse experimental evidence, but results are likely to come in fast, and it may only be a year or two before we can begin to answer what is probably the most important question of this sort: are there any eukaryotes in which split genes are missing?

When we come to consider the actual protein molecules whose genes have been shown to be split, we notice that they all have one thing in common. They are all molecules of terminal differentiation. This is because they are technically the easiest to study. Nobody has yet described or reported an example of a gene for a common-or-garden enzyme (such as one from the Krebs cycle) although such studies are in progress. The other thing that one cannot help noticing is the high frequency of introns. There are two in certain immunoglobulin light chains (*4*), two in various hemoglobin chains (*19–24*), at least four in the γ_1 heavy chain (*29*), and no less than seven in chick ovalbumin (*31*, *32*). Moreover, they are of considerable length, running from just under 100 base pairs to more than 1000. In the ovalbumin gene, the total length of the introns is at least three times that of the exons. If we average this small amount of data, we find that we might expect about one intron for every 300 base pairs or so of exon, and that its average length would be greater than 600 base pairs. That is, on an average, the introns are longer than the exons. In a higher organism a gene has, if anything, more nonsense than sense in it. These preliminary estimates are necessarily very insecure.

The introns in yeast tRNA are much smaller. So far the lengths found are 14, 18, 19, and 34 bases (*41*). Whether there are introns in yeast mRNA is not yet known.

Are there any proteins for which we can say for certain that their genes are not split? This appears to be the case for the sets of histone genes which have been studied both in a sea urchin (*42*) and in *Drosophila* (*43*). In both species, the genes are repeated many times in a tandem arrangement. Unfortunately, there is reason to suspect that histone genes are not completely typical. They do not have polyadenylate [poly(A)] at the end of their mRNA's, for example, and may be designed to exit quickly from the nucleus. It would obviously be interesting to know whether the histone genes of some mammalian species are split or not.

As time goes on, it will be necessary to firm up the preliminary evidence which shows how the transcript of any particular gene is split. The study, by electron microscopy, of the hybridization of genetic DNA with the related mRNA (or of nucleic acid clones derived from them) needs only small amounts of material and, in careful hands, gives reliable results. Historically, it was this method that first suggested that viral mRNA was not a simple colinear transcript of the viral DNA (*2*). Its resolving power is low, however, as is that of mapping by restriction enzyme digestion. For detailed mapping it is essential to obtain the actual base sequences (*44*).

Details and Generalizations

Let us now consider in more detail the arrangement of introns and exons. The first thing we notice, from the very limited experimental data at present available to us, is that a chromosomal gene only produces a single protein (*45*), whereas a stretch of DNA in a virus may produce more than one protein, depending on which way the primary transcript is spliced (*2*). I adopt the attitude that in most cases this is because viruses are short of DNA and, by various devices, their limited amount of DNA is made to code for more proteins than would otherwise be possible. We can see this even in prokaryotic viruses, such as ϕX174, where the same stretch of DNA can be read in one phase for one protein and in another phase for another protein (*46*). A typical example of a "gene" producing more than one protein is the early T-antigen region found in both SV40 (*47*, *48*) and polyoma (*49*). It now seems certain that at least two proteins are produced by this region, each beginning with about 100 residues having exactly the same amino acid sequence. The remaining parts of their amino acid sequences seem to depend on exactly how the RNA transcript is spliced (*50*). Such cases are of interest because the favorable technical nature of the viral systems make it likely that many details will be worked out by studying them. However, such multiple-choice situations may be rare in true chromosomal genes although, as has already been argued (*3*, *4*), they may be important as transitional stages in evolution. Chromosomes seem to have almost more DNA than they know what to do with. Should a chromosomal gene arise whose transcript was processed to make more than one protein, I would expect that in the course of evolution the gene would be duplicated, one copy subsequently specializing on one of the proteins and the other copy on the other. If this point of view is correct, then one would expect multiple-choice genes to occur only rarely in the chromosomes of eukaryotes (*51*).

The other tentative generalization we can make from the present data is that the order of the exons on the DNA is the same as the order in which they are found in the final mRNA. There does not seem any strong reason why this should always be true. It is possible to devise mechanisms in which the order would sometimes be different. This colinearity of the exons probably reflects some important aspect of the origin of introns or of the splicing process and therefore should not be overlooked. It is, incidentally, not true that introns occur only within the *coding* region of a message since in the case of ovalbumin, for example, one intron is found in the leader region of the mRNA before the coding sequence has started.

How Is Splicing Done?

What is the actual mechanism of splicing? At the moment any ideas must necessarily be largely speculative. One would certainly expect at least one enzyme to be involved, if not several. In the case of the tRNA from yeast, an enzyme activity has been found by two groups, as was mentioned above, although it has still to be purified (*11*, *12*). It is not completely obvious that such a mechanism would require a source of energy since two phosphate ester bonds need to be broken whereas only one (or possibly two) have to be made. On balance, one would suspect that energy might be required if only because the process must be an accurate one. Preliminary evidence indicates that the enzyme appears to need adenosine triphosphate (ATP) (*11*). Not all the different kinds of tRNA molecules found in yeast need to be processed by splicing, but so far the indications are that those that are spliced are processed by one and the same enzyme (*41*). There is still no evidence that this same enzyme can also process precursors of mRNA, and I argue that in any case this is unlikely.

This brings us to one of the major unsolved problems: how many different enzymes are involved in splicing? In other words, are some introns removed by one enzyme and other introns by another en-

zyme? I have been so rash as to say, more than once, that we might expect between 10 and 100 different enzymes; but that was pure guesswork. The number could be as low as two.

There are many other major questions to be answered. How does the enzyme (or enzymes) recognize where to splice? This must obviously be done with great precision since the error of a single base would upset the phase of the subsequent part of the message. Is an intron always removed in one go or does the splicing enzyme sometimes need to take several bites at it? What happens to the intron when it is excised? Is it ever used as mRNA? (So far there is no sign of this.) Is it used for control? Is it produced as a linear single-stranded molecule or is it perhaps sometimes excised as a circle? Circularity might increase the stability of the excised molecule. There is little difficulty in thinking of interesting functions which such a single-stranded circular RNA might perform (*52*). Recent work has given us hints to the answers to some of these questions, as I point out below.

Two groups of investigators have worked out the exact base sequences at the borders between the exons and the introns in the ovalbumin gene (*36*, *37*). With one minor exception, the two results agree completely, a tribute to both the rapidity and the accuracy of the present methods for sequencing DNA. Two generalizations arise from these results. Both groups have found that there is often some repetition of the base sequence near the beginning and near the end of an intron. This raises an interesting point which is perhaps not immediately obvious. Imagine that we have the complete base sequence of the DNA in such regions, plus the corresponding sequence on the mRNA. Then, if there is base repetition, we cannot state unambiguously from these data exactly where the splicing actually occurs (Fig. 2). Splicing could conceivably be done in several different ways, and we would still arrive at the same sequence in the mRNA, although the ends of the excised intron would be slightly different, assuming that it had ends.

The sort of repetition actually found is shown in Fig. 3. It can be seen that there is a basic sequence of five, six, or seven bases to which all these marginal sequences are related, in varying degrees. The other generalization, which was pointed out by Chambon's group, is more striking (*37*, *38*). As was explained above, there is always an ambiguity in deciding exactly where the cuts have taken place in the splicing process. Given this ambiguity it is always possible to choose cutting points that obey the following rule: The base sequence of an intron begins with GU and it ends with AG. This rule (*53*) is true not only for ovalbumin but also for the small intron in the λ light chain of immunoglobulin (*4*, *28*). It also appears to be correct for the two hemoglobin introns (*22*, *24*) and for several cases in SV40 (*54*). So far, there is no published exception for an mRNA although there appears to be one in the immunoglobulin heavy chain (*29*). Such a result cannot be due to chance. The rule, however, is not obeyed at the exon-intron junctions found in the tRNA molecules of yeast (*10*, *11*, *41*). This suggests that there are at least two splicing enzymes, one for mRNA and one for tRNA.

Only a few introns have been sequenced completely. The first was the short intron in the earlier part of the mouse immunoglobulin λ_{II} light chain (*4*) and more recently for a λ_I light chain (*28*). The first globin intron has been sequenced for the β-globin of both mouse and rabbit (*22*) and for the α-globin of mouse (*23*). The second intron has also been completely sequenced for mouse α-globin (*23*), but only partially for mouse and rabbit β-globin (*22*). The sequence of part of the large intron in mouse λ_I light chains has just been reported (*28*), and the sequences of three introns and part of a fourth in a mouse immunoglobulin heavy chain have been obtained (*29*) as have the sequences of three complete introns and two incomplete ones of chick ovalbumin (*38*). More sequences will doubtless be reported shortly.

It is difficult to summarize all these data adequately. None of the sequences consists of highly repetitive simple sequences. In those cases tested, they ap-

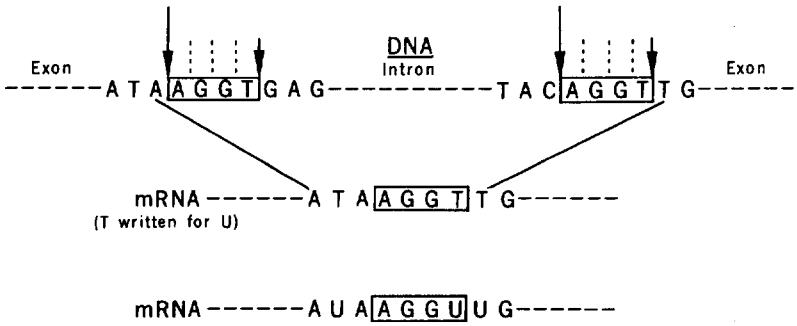

Fig. 2. To demonstrate that the exact cutting position used in splicing may be ambiguous even though the base sequences of both the DNA and the mRNA are known. The top line shows the partial sequence of the DNA of the gene, the bottom line the partial sequence of the resulting mRNA. The middle line is the same as the bottom, but for didactic purposes T (thymine) has been written for U (uracil). It can be seen that the pair of units needed for splicing could either be made where the two big arrows are marked on the top line, or alternatively where the two small ones are located, or at appropriate pairs of positions in between, marked with dotted lines. For a repeat of n bases, $(n + 1)$ pairs of cutting positions are possible. Here $n = 4$. The repeat is shown boxed.

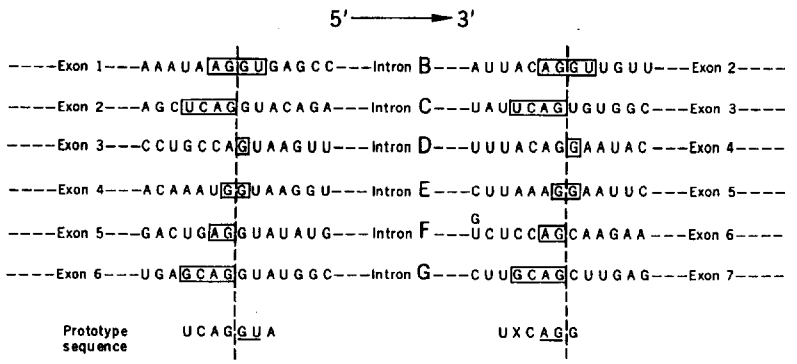

Fig. 3. The base sequences on the RNA primary transcript at the borders between the introns and the exons, deduced from the corresponding DNA sequences. The data is for the ovalbumin gene in chicken and is taken from Breathnach *et al.* (*36*). Catterall *et al.* (*35*) have similar results. Notice that every intron could begin with GU and end with AG. The bottom line shows the "prototype sequences" (*36*). A "concensus sequence," for both junctions (*35*), would be CAGG(U).

pear to be "unique" rather than intermediate repetitive. A number of them tend to be AT rich (T, thymidine), especially T, and not completely random. The 3′ end of an intron sequence often has an unusual base composition, generally T rich. The β-globin introns of mouse and rabbit appear to be distantly related but differ considerably, suggesting considerable drift in evolution (*22*). Only near their margin do the sequences seem to be somewhat more conserved. For ovalbumin there is a suggestion that the sequence is not quite identical in different chickens (*30*, *32*, *33*).

How does the splicing enzyme cut the RNA exactly at the right place? Chambon's rule, noted above, is clearly not enough by itself; nor is the limited amount of base repetition, indicated in Fig. 3, sufficient to select the cutting positions, since similar base sequences occur in other parts of the RNA transcript. The obvious hypothesis is that some secondary or tertiary structure is formed. This would direct the enzyme to the approximate position where cutting is required. Chambon's rule would then allow the enzyme to cut in exactly the right places. This sort of mechanism, a combination of secondary and tertiary structure together with a certain degree of base sequence information, would in any case appear to be plausible on general theoretical grounds (*55*). It remains to be seen whether this hypothetical secondary and tertiary structure can be deduced solely from a study of the base sequences, or whether it will need a direct experimental attack (*56*, *57*).

Other Aspects of RNA Processing

In considering the processing of the primary RNA transcript, it is a mistake to concentrate entirely on the operation of splicing. It may not be true that the extra sequences found in heterogeneous nuclear RNA are due entirely to introns. It also remains to be established whether the operations of RNA trimming (the removal of stretches of RNA at one or both ends of the primary transcript) also takes place and, if so, to what extent. It could conceivably be less important for viral genes than for chromosomal ones.

There is much evidence, admittedly of a rather fragmentary nature, that the primary transcript is packaged in some way on special proteins. It is not at all clear whether this packaging is necessary for successful splicing. (It might, of course, be necessary for mRNA but not for tRNA.) If it is necessary, an interesting possibility arises: are the lengths of the introns quantized in some manner which reflects the way the RNA is combined with the packaging proteins? The present data are perhaps too sparse to permit hazarding a guess on this point.

Most finished mRNA's have a cap at their beginning and a stretch of poly(A) at the end. Recent evidence suggests that, for the late transcript of adenovirus, these terminal additions may occur in the nucleus at a fairly early stage, probably well before splicing takes place (*58*). This would make very good sense. It would not be surprising if these additions to the ends of the RNA molecules in the nucleus protected them from digestion by exonucleases. If this is the case, it is possible to see why splicing has become an important method of processing rather than the trimming, as was envisaged some years ago, in which lengths of RNA were cut off from the beginning or the end of the primary transcript (*59*).

We must also ask what will happen to a transcript containing introns if, for some reason or another, it is not properly spliced. Will it then remain in the nucleus and eventually be degraded? Can the joining part of the splicing mechanism fail, so that after making one or both cuts it leaves the putative mRNA in pieces? Far too little is known about the mechanism by which RNA exits from the nucleus. Can any RNA molecule make its way out? Or are some RNA molecules folded or packaged so that they are unable to penetrate the nuclear pores? Does the presence of a large intron always prevent exiting, simply because its structure is too big to go through a nuclear pore? Is the cap, or something like it, essential for exiting? Does the process require energy? We must consider all these aspects of the steps between the transcription of primary RNA and the appearance of the finished RNA in the cytoplasm.

It is not necessary to assume that the splicing always takes place in the nucleus. It is already known that those tRNA molecules in yeast which need to be spliced are inactive in the unspliced state, both in taking up an amino acid and in functioning on the ribosome (*11*). There is thus no strong reason why they should not first exit from the nucleus, especially as their introns are rather small, and then be spliced in the cytoplasm. However, very recent evidence (*60*) suggests that the enzyme occurs only in the nucleus.

This behavior of the unspliced tRNA molecule is almost certainly a reflection of some feature of its secondary and tertiary structure. It is possible that this may also be true for mRNA although, in general, one would not expect this to form too tight a tertiary structure. The main requirement would seem to be that a ribosome should not be "asked" to attempt to translate a message up to the position where it contains an intron that has not yet been removed. The obvious way is to keep the unspliced transcript within the nucleus until all the splicing is done, but, as we have seen, this may not be true in all cases. In particular, the mRNA transcripts in mitochondria may have evolved in such a way that ribosomes are unable to bind until all the splicing operations have been completed (*61*).

This brings us to the general question of the timing of the splicing process. Does splicing start before the entire transcript is complete? This might seem a sensible thing to do, but preliminary evidence for the late transcripts of adenovirus might suggest that the whole molecule is transcribed before splicing starts (*62*). It is possible that splicing is rather a slow process and does not get under way until the transcript has been finished, or perhaps that there are special mechanisms to prevent premature splicing. Clearly there are many complicated experimental questions which remain to be answered.

Evolution of Splicing

It is impossible to think about splicing for long without asking what it is all for. In particular, what would happen to the functioning of a gene if a particular intron were removed completely? This leads us to ask how splicing arose in evolution. I have noticed that this question has an extraordinary fascination for almost everybody concerned with the problem. It might be thought rash to inquire too closely about the origins of a mechanism when we do not yet know exactly how it works at the present day. This gap in our knowledge does not deter speculation, and for good reason, for such speculation may suggest interesting ideas and perhaps give us some general insight into the whole process. Unfortunately, there is a tendency to fall into the fallacy of evolutionary foresight. For a change in a genome to spread through the population it must usually have a selected advantage, although occasionally it may spread by "hitchhiking" on the selective advantage of an associated part of the genome. Even if it has already spread, it cannot remain indefinitely without having some advantage since otherwise it will eventually be deleted.

Thus, one should not invoke some selective advantage occurring only in the future unless this is likely to happen within a time comparable to the time needed to remove the intron (*63*).

This problem should not be confused with the related phenomenon of a particular stretch of DNA spreading within one genome, the case of "selfish DNA." The advantage such DNA needs is simply that, by one mechanism or another, it replicates during evolution rather more than the bulk of the DNA and that, in doing so, it does not do too much harm to its "host." Any complete discussion of the evolution of eukaryotic genomes must take into account such preferential replicators (*64*).

With these reservations in mind let us, nevertheless, attempt to paint some broad evolutionary picture. The first problem is that of timing. When did introns first arise? The obvious suggestion is that they came in with the eukaryotes. Two investigators (*65*, *66*) have proposed they originated at a much earlier time. This issue may prove difficult to resolve, and I shall not pursue it further here. Nor shall I discuss the possible origin of the splicing enzymes.

Three possible mechanisms have been suggested for the formation of a new intron. To make the discussion simpler, I shall assume that the splicing signals on the RNA lie mainly near the boundaries of an intron, although the real situation is likely to be more complex.

1) The splicing signals arise, accidentally, in a stretch of DNA which is already being transcribed (*29*). For the very first intron, the signals that the first splicing enzyme happened to recognize could have already been in existence. Those for later introns would have had to arise by random mutation. Thus, a portion of the RNA transcript becomes spliced out so that the mRNA and the protein it codes for are both shortened. The base sequence of the intron, no longer used for coding, then drifts rather rapidly. This idea can be extended to cover other, similar situations.

2) An intron is inserted in the middle of a piece of DNA by a special insertion mechanism (*67*) that automatically generates flanking sequences closely related to those required, on the RNA transcript, for splicing (*68*). Then few mutations, if any, are required to initiate some degree of splicing.

3) A new intron is produced by translocating an exon (by any mechanism) together with parts of its flanking introns (*3*, *4*). For example, this DNA might be inserted into an already existing intron, thus producing two introns where there was only one before. This process also might automatically generate the splicing sequences required for the new introns, although further mutations might be needed to make splicing efficient.

This last idea, that of exon shuffling, was first advocated by Gilbert and Tonegawa (*3*, *4*). It has at least two advantages. New proteins can be produced by bringing together amino acid sequences that have already been evolved separately to fold up neatly and to perform some function or other, rather than by adding lengths of "random" base sequences to an existing protein (*69*). The mechanisms for selecting these DNA sequences need not be very precise since the edges of the insert could be located almost anywhere in the flanking introns and, if it were inserted into an intron, the exact position of insertion would not matter either.

On this theory, then, the DNA sequence coding for globin did not start in evolution as a single uninterrupted stretch of DNA. Instead, it evolved from three distinct exons, which already existed, and which random shuffling brought together in the genome. The resulting two introns on the RNA were spliced out to produce, for the first time, the typical globin sequence. Gilbert has pointed out to me that the middle exon in globin codes for that part of the polypeptide chain which embraces the heme group. It could well have been taken from a heme-containing protein, which had evolved earlier. Whether the present first and third exons at one time coded for a single protein is also a matter for speculation.

Some of the present evidence supports this theory. The great age of the two globin introns (*23*) and the fact that no other introns have so far been found there show that the successful production of new introns is probably a rare event. The position of introns in the immunoglobulins separating the structural domains of the proteins, and the intron toward the end of the signal peptide are just what one might expect. A signal peptide is exactly the sort of amino acid sequence it would be useful to shuffle around since its addition to a cytoplasmic protein could convert it into an excreted one (*70*).

Which of these three ideas is correct? At the moment it is impossible to tell, especially as they may all have contributed, at one time or another, to the production of new introns. The first mechanism, that of random mutation, would appear to be rather a rare event, but it might have been enough to get splicing started. The second mechanism, that of specific insertion, could well have produced the first introns. Perhaps splicing evolved as a defense by the cell against an insertion element it was harboring. The fact that no globin genes have yet been found to have a third intron suggests that if this mechanism still operates in evolution it does so at a fairly low rate.

The third mechanism, that of exon shuffling, looks like a very plausible explanation for the origin of those introns found between protein domains. It is obviously a likely mechanism in those organisms which have a large intron-exon ratio, but it would probably work less well in one with few, rather small introns, if such organisms exist. Thus a reasonable guess, as supposed by Tonegawa (*29*), might be that introns first originated by one of the first two mechanisms, but that organisms with a large percentage of intron DNA have produced most of their more recent introns by exon shuffling. However, both Doolittle (*65*) and Darnell (*66*) feel that exon shuffling is so advantageous for evolution that they believe it originated at a very early stage.

How easy is it to delete an existing intron completely? If an intron has acquired some essential function, its deletion will be selected against, but suppose it has little or no function. Even in such a case, the deletion of an entire intron may be a rare event since it must be done very precisely to produce a functioning mRNA. What one might expect is that random deletions could continually reduce the size of such an intron. However, the continual shuffling around of DNA in evolution probably adds DNA to introns in a rather random manner. Thus, the length of an intron may represent a dynamic balance in evolution between additions and deletions.

If introns are indeed difficult to remove, it can be seen that once a sufficient number have been introduced it would be impossible to delete the splicing enzyme (or enzymes) without catastrophic consequences for the organism. For the same reason the specificity of the splicing enzyme (or enzymes) is likely to be very similar in many different species. It should be almost impossible to get rid of a splicing enzyme except under very heavy evolutionary pressure. Doolittle (*65*) has already suggested that this is what happened in most prokaryotes.

Control of RNA Processing

Once introns are common, it is more than likely that evolution would eventually start using them for other purposes such as control. It seems almost certain that there will be some control of gene expression at the transcriptional level,

but this does not mean that there may not be additional controls at the processing level (*71*). In rather general terms, we can conceive of this as being of two types. The first would be a rather coarse control by which large groups of genes were switched on and off simultaneously. This could come about if there were several different enzymes for splicing. If, at some stage in differentiation, one of these were absent, then all those transcripts that required it would not be able to form functional mRNA. This could clearly be a useful control for major steps in the developmental process. Unfortunately, if Chambon's rule is true, it hints that there may be only one enzyme for mRNA. This tentative conclusion could easily be incorrect. For example, the splicing enzymes may perhaps be in two parts, one of which is always the same and performs the recognition of the GU and AG base sequences and the actual cutting operation, while the second part recognizes some other feature of the base sequence or of the secondary or tertiary structure. Only the purification of the splicing enzymes will prove this point.

It is also possible to imagine a fine control that might apply only to a single intron or to a small number of them. As opposed to the coarse control, which would be a positive one, this might be a negative control (*72*). This hypothetical repressor protein would combine in some specific way with a particular intron so that the splicing enzyme was unable to operate. It seems to me more than likely that nature will have evolved such a process for some introns, but I should be very reluctant to guess just how many of them might be controlled in this particular way.

Shuffling of Controls

If the insertion or translocation of DNA does occur in evolution—and indeed there is much indirect evidence that something of this sort is taking place—will these additions be made in special places in the DNA or will they go in more or less at random? If they go at random, or at base sequences which occur fairly frequently, we should certainly expect them to be put into those regions that are not transcribed, including those that function for the control of the transcription. Indeed, it has often been argued that this is exactly where we need more sequences in higher organisms since the evolution of complex cellular organisms may require more intricate and flexible control mechanisms. Perhaps the main selective advantage for insertions will come from those put into the noncoding regions. This might imply that some of the insertions we find today within those portions of the DNA coding for a single polypeptide chain are merely an accidental and often unnecessary byproduct of a process whose main function is to evolve more subtle patterns for the control of transcription.

To grasp what has been happening in evolution we shall have to understand all the mechanisms by which stretches of DNA can be multiplied in the genome or added to or subtracted from it. These would include the possible jumping or jittering of DNA polymerase, recombination events of all types (especially for tandem repeats), deletion mechanisms, insertions due to viruses and other replicating entities, transformation and various translocation mechanisms, both specific and nonspecific. The theory of the "selfish gene" will have to be extended to any stretch of DNA. A molecular biologist who wishes to discuss the evolution of the eukaryotic genome will need not only to know a lot about the way DNA and its transcripts can behave but also something about modern ideas on population genetics.

Nucleic Acid Taxonomy

This naturally brings one to the taxonomic implications of introns and insertions, wherever they may appear in the genome. We can confidently predict that there will be an enormous expansion in our knowledge of all types of sequences, not only exons and introns and the regions adjacent to them, but also of repetitive sequences and simple sequences of all kinds. People interested in molecular taxonomy are going to have a field day. It is virtually certain that discoveries will turn up which will radically alter our ideas of the details of the evolutionary process (*73*). I would not be surprised if the base sequence of large parts of the introns drifted at a fairly rapid rate; there is already some evidence for this (*20*, *22*, *23*). If so, such sequences would be excellent tools with which to study the shorter periods of evolution. By contrast, the introduction of completely new introns may occur only rarely, and their study may be useful for much longer periods in the evolution process. If an intron is sometimes altered in length, as is the larger intron in the β-globin gene (*21*), such a change would also provide a useful evolutionary marker, possibly on an intermediate time scale. The recent advances in DNA splicing, together with the new and rapid methods of sequencing DNA, have made it entirely possible for such studies to be conducted on many different genes in many different individuals in many different species.

Conclusion

There can be no denying that the discovery of splicing has given our ideas a good shake. It was of course already surmised that the primary RNA transcript would be processed in some way, but I do not share the view sometimes expressed that splicing is only a trivial extension of our previous ideas. I think that splicing will not only open up the whole topic of RNA processing, which had become somewhat bogged down before splicing was discovered, but in addition will lead to new insights both in embryology and in evolution. What is remarkable is that the possibility of splicing had not at any time been seriously considered before it was forced upon us by the experimental facts. This was probably because, looking back, we can see that there was no earlier experimental evidence to suggest that such a process might be taking place, at least for mRNA. Lacking evidence we had become overconfident in the generality of some of our basic ideas.

Splicing then, in spite of the patchiness of the evidence, is almost certainly a real process and probably an important one. Further studies on the base sequences involved and, in particular, on the enzymes performing the operation are likely to increase our knowledge of it fairly rapidly. Before long one might hope to understand all the various processing steps, the trimming (if it exists), the capping, the packaging, the addition of poly(A), the splicing, and the exit from the nucleus, if not in full detail, then at least in outline. But our enthusiasm for this exciting new field should not let us lose sight of the even more fundamental process preceding it: transcription and the control of transcription. Here we badly need additional breakthroughs, both experimental and conceptual, before we can feel we have a real grasp of gene structure, gene control, and gene evolution in eukaryotes.

References and Notes

1. Throughout the article I have deliberately used the word "gene" in a loose sense since at this time any precise definition would be premature.
2. *Cold Spring Harbor Symposium on Quantitative Biology* (1977), vol. 48.
3. W. Gilbert, *Nature (London)* **271**, 501 (1978).
4. S. Tonegawa, A. M. Maxam, R. Tizard, O. Bernard, W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1485 (1978).

5. There are two main difficulties. A stretch of nuclear RNA may be part of an intron if spliced one way but part of an exon if spliced in another way. In addition, one is tempted to use the two words to describe the parts of the DNA from which the stretches of RNA are transcribed. Nevertheless, used judiciously, the two words are undoubtedly useful. I imagine some committee will eventually decide on a wholly logical terminology.
6. N. Hozumi and S. Tonegawa, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 3628 (1976); C. Brack and S. Tonegawa, *ibid.* **74**, 5652 (1977); T. H. Rabbitts and A. Forster, *Cell* **13**, 319 (1978).
7. C. Brack, M. Hirama, R. Lenhard-Schuller, S. Tonegawa, *Cell* **15**, 1 (1978); J. G. Seidman and P. Leder, *Nature (London)* **276**, 790 (1978).
8. The second seems unlikely for large introns if only because RNA polymerase will be transcribing not from naked DNA but from chromatin. The third implies a very specific bimolecular reaction. The low concentrations of the components in the nucleus would make this a somewhat slow process.
9. H. M. Goodman, M. V. Olson, B. D. Hall, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5453 (1977).
10. P. Valenzuela, A. Venegas, F. Weinberg, R. Bishop, W. J. Rutter, *ibid.* **75**, 190 (1978).
11. P. Z. O'Farrell, B. Cordell, P. Valenzuela, W. J. Rutter, H. M. Goodman, *Nature (London)* **274**, 438 (1978).
12. G. Knapp, J. S. Beckmann, P. F. Johnson, S. A. Fuhrman, J. Abelson, *Cell* **14**, 221 (1978).
13. J. L. Bos, C. Heyting, P. Borst, *Nature (London)* **275**, 336 (1978).
14. D. M. Glover and D. S. Hogness, *Cell* **10**, 167 (1977); R. L. White and D. S. Hogness, *ibid.*, p. 177; P. K. Wellauer and I. B. Dawid, *ibid.*, p. 193; M. Pellegrini and J. Manning, *ibid.*, p. 213; P. K. Wellauer, I. B. Dawid, K. D. Tartoff, *ibid.* **14**, 269 (1978).
15. I. B. Dawid, personal communication.
16. R. A. Krzyzek, M. S. Collett, A. F. Lau, M. L. Perdue, J. P. Leis, A. J. Faras, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1284 (1978).
17. P. P. Slonimski, M. L. Claisse, M. Foucher, C. Jacq, A. Kochko, A. Lamouroux, P. Pajot, G. Perrodin, A. Spyridakis, M. L. Wambier-Kluppel, in *Biochemistry and Genetics of Yeast, netics of Yeast*, M. Bacila, B. L. Horecker, A. O. M. Stoppani, Eds. (Academic Press, New York, in press); P. Borst *et al.*, in *Mitochondria 1977: Genetics and Biogenesis of Mitochondria*, W. Bandlow, *et al.*, Eds. (De Gruyter, Berlin, 1977), pp. 213–254.
18. G. Bernardi, *Nature (London)* **276**, 558 (1978); P. Borst and L. A. Grivell, *Cell* **15**, 705 (1978).
19. A. J. Jeffreys and R. A. Flavell, *Cell* **12**, 429, 1097 (1977); S. M. Tilghman, D. C. Tiemeier, J. G. Seidman, B. M. Peterlin, M. Sullivan, J. V. Maizel, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 725 (1978); S. M. Tilghman, P. J. Curtis, D. C. Tiemeier, P. Leder, C. Weissmann, *ibid.*, p. 1309; A. J. Kinniburgh, J. E. Mertz, J. Ross, *Cell* **14**, 681 (1978); R. A. Flavell, J. M. Kooker, E. De Boer, P. F. R. Little, R. Williamson, *ibid.* **15**, 25 (1978); J. G. Mears, F. Ramirez, D. Leibowitz, A. Bank, *ibid.*, p. 15.
20. D. C. Tiemeier, S. M. Tilghman, F. I. Polsky, J. G. Seidman, A. Leder, M. H. Edgell, P. Leder, *Cell* **14**, 237 (1978).
21. R. M. Lawn, E. F. Fritsch, R. C. Parker, G. Blake, T. Maniatis, *ibid.* **15**, 1157 (1978).
22. J. van den Berg, A. van Ooyen, N. Martei, A. Schamböck, G. Grosveld, R. A. Flavell, C. Weissmann, *Nature (London)* **276**, 37 (1978).
23. A. Leder, M. Miller, D. Hamer, J. G. Seidman, B. Norman, M. Sullivan, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 6187 (1978).
24. D. A. Konkel, S. M. Tilghman, P. Leder, *Cell* **15**, 1125 (1978).
25. H. I. Miller, D. A. Konkel, P. Leder, *Nature (London)* **275**, 722 (1978).
26. T. H. Rabbitts, *ibid.*, p. 291.
27. J. G. Seidman, A. Leder, M. H. Edgell, F. Polsky, S. M. Tilghman, D. C. Tiemeier, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 3881 (1978).
28. O. Bernard, N. Mozumi, S. Tonegawa, *Cell* **15**, 1133 (1978).
29. S. Tonegawa, personal communication.
30. R. Weinstock, R. Sweet, M. Weiss, H. Cedar, R. Axel, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1299 (1978).
31. A. Dugaiczyk, S. L. C. Woo, E. C. Lai, M. L. Mace, Jr., L. McReynolds, B. W. O'Malley, *Nature (London)* **274**, 328 (1978).
32. J. L. Mandel, R. Breathnach, P. Gerlinger, M. Le Meur, F. Gannon, P. Chambon, *Cell* **14**, 641 (1978).
33. E. C. Lai, S. L. C. Woo, A. Dugaiczyk, B. W. O'Malley, *ibid.* **16**, 201 (1979).
34. S. L. C. Woo, A. Dugaiczyk, M. Tsai, E. C. Lai, J. F. Catterall, B. W. O'Malley, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 3688 (1978).
35. J. F. Catterall, B. W. O'Malley, M. A. Robertson, R. Staden, Y. Tanaka, G. G. Brownlee, *Nature (London)* **275**, 510 (1978).
36. R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, P. Chambon, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4853 (1978).
37. P. Kourilsky and P. Chambon, *Trends Biochem. Sci.* **3**, 244 (1978).
38. M. A. Robertson, R. Staden, Y. Tanaka, J. F. Catterall, B. W. O'Malley, G. G. Brownlee, *Nature (London)*, in press.
39. J. L. Nordstrom, D. R. Roop, M. J. Tsai, B. W. O'Malley, in preparation.
40. Y. Suzuki, as reported [in *Nature (London)* **275**, 364 (1978)] by W. J. Gehring.
41. J. Abelson, personal communication.
42. W. Schaffner, G. Kunz, H. Daetwyler, J. Telford, H. O. Smith, M. L. Birnstiel, *Cell* **14**, 655 (1978).
43. D. S. Hogness, personal communication.
44. It is also important to make certain that the mRNA being studied is in fact copied from the DNA being studied and not from some closely related but silent DNA sequence. This is best established in most cases by the detailed sequencing of a mutant version of the gene. This has already been done for one of the tRNA genes of yeast (*9, 11*).
45. The immunoglobulin chains may be an exception.
46. F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, M. Smith, *Nature (London)* **265**, 687 (1977); F. Sanger, A. R. Coulson, T. Friedman, G. M. Air, B. G. Barrell, N. L. Brown, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, M. Smith, *J. Mol. Biol.*, in press.
47. L. V. Crawford, C. N. Cole, A. E. Smith, E. Paucha, P. Tegtmeyer, K. Rundell, P. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 117 (1978).
48. W. Fiers, R. Contreras, G. Haegeman, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert, M. Ysabaert, *Nature (London)* **273**, 113 (1978); V. B. Reddy, B. Thimmappaya, R. Dhar, K. N. Subramanian, B. S. Zain, J. Pan, P. K. Ghosh, M. L. Celma, S. M. Weissman, *Science* **200**, 494 (1978).
49. M. A. Hutchinson, T. Hunter, W. Eckhart, *Cell* **15**, 65 (1978).
50. Paradoxically the shortest protein is produced by the longest mRNA because, after about 174 codons, a region with several chain terminators occurs (*48*). This region is spliced out in the other messengers, with the result that longer polypeptide chains are possible.
51. So far there is nothing to suggest that the relative abundances of the different protein products of a multiple choice viral gene are different in different circumstances. That is, there is as yet no evidence for a control mechanism that adjusts the relative abundance of the members of such a protein set, although I see no reason why this should not happen in some cases.
52. So far, the only circular single-stranded RNA's known are the plant viroids [H. L. Gross *et al.*, *Nature (London)* **273**, 203 (1978)]. How these are related to introns is still unknown. A small amount of circular RNA in the nucleus could easily have been overlooked.
53. The actual data suggest a slightly stronger rule for the first cut: that it is made between the two G's in GGU (G, guanine; U, uracil). There is one exception to this in which the sequence is AGU (A, adenine). Thus, so far the first cut could always be before G in the sequence RGU, where R is a purine. In almost all cases, both cuts might be one position earlier; that is, before the first G of GGU, at the beginning of the intron, and before the G of AG at its end. It will be necessary to sequence the ends of the encised introns in each case to decide exactly where the cuts are made.
54. P. K. Ghosh, V. B. Reddy, J. Swinscoe, P. Lebowitz, S. M. Weissman, *J. Mol. Biol.*, in press.
55. If the splicing positions were controlled solely by a unique base sequence that was always the same, there might be a danger that wrong splices would occur (if a gene had more than one intron) since the mechanism might accidentally splice from the beginning of one intron to the end of another one.
56. It will be recalled that for tRNA the first method easily yielded the secondary structure but, in spite of brave attempts, the tertiary structure proved too difficult and had to be obtained by x-ray diffraction studies on crystals of tRNA.
57. A naïve application of Chambon's rule suggests that in some cases the cut might be made either way round. That is, the intron might occasionally be joined to form a circle, leaving the mRNA with a gap in it. More plausibly, the enzyme might splice the exons together while at the same time making the intron circular. An example of such a reciprocal sequence is the first one listed in Fig. 3.
58. E. B. Ziff and R. M. Evans, *Cell* **15**, 1463 (1978).
59. If a cap and a poly(A) tail are necessary for most messengers in the cytoplasm, either to give them stability or, in the case of the cap, to assist in the ribosomal binding process; and if these additions are made in the nucleus fairly soon after the relevant parts of the transcript became available in order to give the transcript stability, then splicing would be the only remaining operation open to the transcript.
60. R. D. Kornberg, personal communication.
61. It will be interesting to see, if splicing does occur in mitochondria, whether the enzyme (or enzymes) used is related to one or other of the splicing enzymes in the rest of the cell and, if so, to which.
62. See the review by J. E. Darnell, Jr., to appear in *Prog. Nucleic Acid Res.* and *Mol. Biol.* There is also evidence (not documented here) for long transcripts of the globin, ovalbumin, and ovomucoid genes.
63. Not all inserts now present need have a function. For all we know a fair proportion of them may be sitting there, doing nothing, and simply waiting to be excised or deleted.
64. R. Dawkins, *Z. Tierpsychol.* **47**, 61 (1978). I am indebted to L. E. Orgel for making this point.
65. W. F. Doolittle, *Nature (London)* **272**, 581 (1978).
66. J. E. Darnell, Jr., *Science* **202**, 1257 (1978).
67. M. P. Calos and L. Johnsrud, *Cell* **13**, 411 (1978); N. D. F. Grindley, *ibid.* p. 419. For a general reference see *DNA, Insertion Elements, Plasmids, and Episomes*, A. I. Bukhari, J. A. Shapiro, S. L. Adhya, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1977).
68. F. H. C. Crick, *Eur. J. Biochem.* **83**, 1 (1978).
69. For a discussion of this idea see C. C. F. Blake [*Nature (London)* **273**, 267 (1978)].
70. Another interesting case is that of the large precursor protein that contains the amino acid sequences for ACTH, β-MSH, β-lipotropin (which itself contains the endorphins) [R. E. Mains, B. A. Eipper, N. Ling, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 3014 (1977)] and possibly other hormones. It will be interesting to see, when that particular gene is sequenced, whether there are introns between the hormone sequences.
71. For another discussion of this topic see (*62*).
72. It would resemble the action of the lac repressor in interfering with the function of RNA polymerase, for example.
73. There are already hints that there may be even more genetic polymorphism in wild populations than that already established by the study of amino acid sequences.
74. References have been kept to a bare minimum. I hope those working on mammalian viruses and the immune system will forgive me for not describing their results more fully. To do so would have made the article far too long.

I thank J. Abelson, G. G. Brownlee, P. Chambon, J. E. Darnell, Jr., I. B. Dawid, D. S. Hogness, P. Leder, B. W. O'Malley, P. P. Slonimski, S. Tonegawa, S. M. Weissman, and E. B. Ziff for providing me with unpublished material, and J. Abelson, P. Chambon, W. Gilbert, L. E. Orgel, and S. Tonegawa for useful comments on the manuscript. This work was supported by the Eugene and Estelle Ferkhauf Foundation, J. W. Kieckhefer Foundation, and Samuel Roberts Noble Foundation, Inc.