Biggest Challenge Since the Double Helix

Recombinant DNA technology is beginning to make a significant impact on our knowledge of gene expression in higher organisms

The discovery in the early 1970's of enzymes that cut and splice DNA in a controlled fashion led initially to a burst of gene manipulation motivated almost as much by the fact that it could be done as by what could be done with it. "Cloning fever," as Brian McCarthy of the University of California at Irvine describes it, has passed, and "people are beginning to ask some really interesting biological questions." These questions, relating to the structure and function of genes from higher organisms, were outlined at the First Annual Congress for Recombinant DNA Research held recently in San Francisco.

How are the genes of higher organisms expressed: that is, how is the genetic information encoded in the DNA first transcribed into an RNA copy and then translated into the amino acid sequence of a protein? And how are the myriad genes of a cell coordinated so that they function as an integrated part of an organism? These are the targets for the increasingly sophisticated techniques of recombinant DNA technology.

The study of gene expression falls into four major areas: first, gene structure (that is, the precise sequence of nucleotide bases that make up DNA) and the elements that initiate expression; second, the influence of the way in which the DNA molecule is packaged in the chromosome; third, the steps through which the RNA message passes between its transcription from the DNA template and its exit from the nucleus to the cytoplasm; fourth, the translation of genetic information in the message into the sequence of amino acid in a protein. The tools of recombinant DNA technology are being applied primarily to the first three.

Bacteria, particularly the common gut bacterium *Escherichia coli*, have been the prime targets for molecular biological analysis for several decades. With relatively few genes, a chromosome that isn't secreted away in a nucleus, and no specialization into different cell types that are found in higher organisms, bacteria represented relatively tractable experimental models. The result is that *E. coli* is now the most thoroughly understood organism in the world. With the advent of recombinant DNA technology, study of the more complex nucleated cells of higher organisms at last became a manageable prospect. The question was, would the pattern of gene structure and the transcription initiation sequences worked out in bacteria be repeated in higher cells? The answers are: with the initiation switch, up to a point; and with the gene structure, most definitely not.

A gene is transcribed when RNA polymerase, a massive enzyme, runs along the length of DNA, copying the sequence of nucleotide bases [a combination of adenine (A), thymine (T), guanine (G), and cytosine (C)] into their RNA equivalent. The first step in the process is the specific binding of the enzyme to a precise location at the beginning, or 5′ end, of the gene. This "promoter" region determines the site at which initiation begins and the efficiency with which it is carried out. The search for the bacterial promoter occupied a good deal of research effort, which eventually paid off. As predicted, the structure of the promoter from nucleated cells is proving more elusive, but within the past few months significant progress has been made. Interest in the promoter is, incidentally, more than academic, since a switch that can efficiently turn genes on would be extremely valuable in the world of commercial gene manipulation.

> The biggest surprise in recent years is that genes of nucleated cells are fragmented. . . . A significant number are strung out in more than 20 pieces, and the record is a gene in 52 pieces.

With surgical precision, recombinant DNA researchers have been slicing away nucleotides piece by piece in the region immediately in front of a number of genes so as to discover how much of that section is required for correct initiation of transcription. "Two regions are especially important," says Pierre Chambon of Université Louis Pasteur, Strasbourg. "One is around 30 nucleotide bases upstream from the site of transcription initiation, and the other is around 80 bases upstream," he says. The first of these, at position denoted −30, is known as the TATA box, because of the double combination of adenine and thymine nucleotide bases. "A year ago the evidence for the importance of the TATA box was circumstantial," says McCarthy, who opened the first session of the San Francisco meeting. "But Chambon's experiments make it clear that it determines the exact site of initiation."

The TATA box does not operate in isolation. A set of nucleotide sequences of the pattern CCAT (the CAT box) at position −80 also exerts an influence on the efficiency of transcription. And there are hints that another region at position −150, and others even further upstream, may be important too.

The promoter for nucleated cells may well be in sight, and it does bear striking similarities with the bacterial promoter. But it is also clear from experiments described at the San Francisco meeting that genes in higher organisms may be influenced by sites far upstream. "It's going to take a while before all of these sites are located," says Bert O'Malley, of the Baylor College of Medicine, Houston.

The biggest surprise in molecular biology in recent years was the discovery that instead of being arranged as a continuous string of nucleotide bases, the genes of nucleated cells (and viruses that invade the nuclei of such cells) are fragmented. Bacterial genes are not frag-

0036-8075/81/0403-0028$00.75/0 Copyright © 1981 AAAS

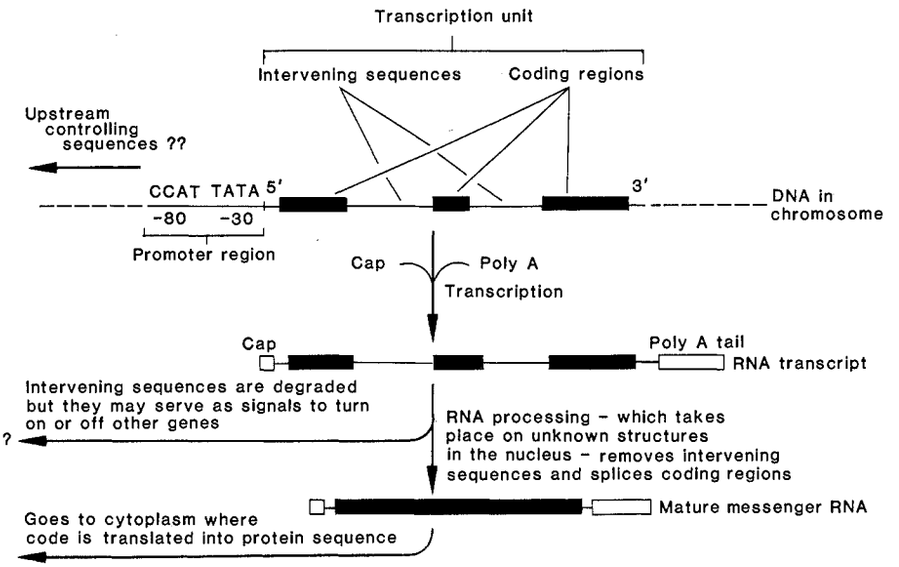
mented. The question initially asked was, how many genes from higher organisms would have this piecemeal structure? All but a tiny handful. It is not just a question of genes being in two or three pieces. A significant number are strung out in more than 20 pieces, and the record so far is held by α-collagen, which is in 52 pieces.

One effect of this extraordinary arrangement is that genes in nucleated cells occupy far longer sections of DNA than do their bacterial equivalents. Often the coding sequences are dwarfed in size by the intervening sequences. So, a gene that might require a combination of 1000 nucleotide bases to code for an average size protein might be spread out over a length of DNA ten times that size, with the promoter region perched at the 5′ end. Although this discovery goes some way toward solving the puzzle of why nucleated cells have so much apparently excess DNA, it raises many questions. Specifically: Why did this arrangement arise? How do cells cope with the mechanics of removing the intervening sequences so as to make correctly structured messenger RNA molecules? And what role might this arrangement play in the day-to-day life of the cell?

The most attractive proposal for why the fractured gene arrangement became established in nucleated cells concerns its benefits to evolutionary change. It is becoming clear that, through a number of mechanisms, DNA in chromosomes is subject to a surprising degree of movement: sections of DNA may be excised from one location and inserted in another, or merely be swapped from one chromosome to its partner chromosome. Such rearrangements may be harmful if they disturb existing genes, neutral if they involve a stretch of DNA with no genes, or beneficial if they create a new functional gene.

If the fragmented arrangement seen in genes from higher organisms represents a combination of functional units, that is, if each coding sequence specifies an area that does a specific job in the final protein, then shuffling pieces of DNA around the chromosome is potentially extremely beneficial to the organism. Several functional units might come to lie in tandem, creating a new gene, which has the now familiar split gene structure.

When Walter Gilbert, of Harvard, first put forward this idea, there was little experimental evidence to support it. To be sure, antibody molecules appear to fit this picture perfectly. They are proteins with a large number of functional units, and each unit is coded for by a distinct coding region of highly fragmented



Stages in expression of genes from higher organisms

Genetic information encoded in the DNA of the chromosomes is first transcribed into an RNA copy which then forms the template for translation into the amino acid sequence of a protein. In contrast with bacterial genes, most genes in higher organisms are fragmented as discrete coding regions separated by noncoding intervening sequences. As transcription produces a faithful copy of the mosaic of coding and noncoding sequences, the RNA message has to be edited before it passes into the cytoplasm: this involves excision of the intervening sequences followed by precise splicing of the coding regions. Control of initiation of transcription and the mechanism of splicing are two major puzzles of contemporary molecular biology.

genes. But it can be argued that antibodies are a special case: they require an extensive combinatorial ability in order to produce a large selection of molecules with different specificities.

There are other examples that are not special cases, but not many. Although it wasn't obvious at first, the way the globin gene is fragmented corresponds with distinct functional units in the protein molecule. O'Malley and his colleagues have just published data on the ovomucoid protein, the functional units of which are echoed in the fragmented structure of its gene. Biologists are surprisingly naïve about the functional fine structure of most protein molecules, but can it be possible, for instance, that conalbumin has 17 functional units and ovalbumin 8?

In his work with human interferon Charles Weissmann, of the University of Zurich, mimicked the potential benefits of shuffling gene fragments. There are a number of different interferon molecules, each produced by a slightly different gene. Weissmann showed that by combining fragments from different genes, thus producing a hybrid interferon gene, he could create new interferons, some of which were significantly more effective in biological tests than the natural antiviral proteins.

Evolutionary payoff may well have established the fragmented structure of genes from higher organisms, and future evolution might indeed make use of this arrangement. But it is reasonable to suppose that the intervening segments have assumed a role in the daily lives of cells, otherwise they would surely have been deleted through natural selection.

Genes for the most part do not work in isolation: they may be part of a complex temporally controlled system. Imagine, then, when a gene is transcribed: The transcript is a faithful copy of the gene, with its mosaic of coding and noncoding regions. The noncoding, intervening, sequences are then cut out, and the coding sequences are spliced, and the full intact messenger goes on its way to the cytoplasm where protein synthesis takes place. Could it be that the excised intervening sequences act as signals to other genes, saying "gene A has just been transcribed; it's gene B's turn to be expressed"? Coordination of gene action has to be accomplished by some means, and this system has potential. It is an attractive idea, with no experimental support as yet.

There is, however, one recently discovered example of an intervening sequence coding for an enzyme that promotes the excision of that sequence and the further maturation of the mosaic transcript. This is the gene for cytochrome b in yeast mitochondria. How general such self-regulation is remains to be seen.

The fragmented gene structure offers a

second possible contribution to the daily life of an organism, through variable splicing. If the excision of intervening regions and subsequent joining of coding regions can take place in, say, two slightly different patterns, then a single gene could give rise to different messenger RNA's and therefore to two different protein molecules. "We appear to have an example of this with human growth hormone," says John Baxter of the University of California, San Francisco. "One of the intervening sequences in the human growth hormone transcript is processed in two ways, giving two hormone variants. One of them has a molecular weight of 20,000 and growth stimulating properties but does not affect carbohydrate metabolism; the other is 10 percent larger and has both metabolic properties." Variable splicing is clearly a possible fertile source of metabolic nuance.

When molecular biologists talk about their attempts to analyze gene action, it is easy to get the impression that DNA is naked in the nucleus, that all that is important is the sequence of nucleotide bases that make up the DNA strand. DNA is not naked in the nucleus: it is packaged by being coiled around tiny protein spheres, producing the effect of a string of beads. The combination of DNA and protein beads is known as chromatin and is special to nucleated

> "Chromatin must have a profound effect on the way genes work in these cells."

cells. "Chromatin must have a profound effect on the way genes work in these cells," emphasizes O'Malley. The question is, is the effect a passive one arising from the physical constraints of packaging, or does it involve more specific interactions that determine that one gene shall be active in a particular cell while another remains silent?

When a cell differentiates, that is, becomes specialized as, for instance, a liver cell, a particular suite of genes pertinent to the cell's future activity is put into a state of readiness. "This involves a different conformation of chromatin," explains O'Malley. "The genes, and a large area around them, are now sensitive to a DNA digesting enzyme." Once a gene receives a signal to start transcribing, its sensitivity to DNA digestion increases still further, especially around the initiation site. This implies that the arrangement of chromatin beads is different—more open—in pregnant and active genes, and some researchers claim that this rearrangement derives from specific interactions between the DNA sequences and the proteins. Others dispute the specificity of interaction. This is an area where the meeting of molecular and cellular biology has yet to yield unequivocal results.

One specific suggestion for DNA/protein interaction in chromatin relates to the structure of the promoter region. Harold Weintraub, of the Fred Hutchinson Cancer Center, Seattle, suggests that packaging in this area aligns functional sequences in important ways. "One turn of the DNA strand round a bead would bring the TATA box and the CAT box right next to each other," he claims. "And another turn would align them with the next upstream region at −150." This comment may display a keen eye for the importance of three-dimensional structure in the interaction of DNA with the RNA polymerase. Or the spacing of these regions may be merely coincidental. It turns out that at least some genes are devoid of protein beads around the promoter region, so that the coiling notion cannot apply in these cases. Currently the alignment of sites on beads is interesting speculation.

Just as DNA does not go naked in the nucleus, so too does RNA find itself clothed in protein molecules during its journey from the site of transcription to the pore in the nuclear membrane through which it exits. This stage of gene expression—termed RNA processing—has more puzzles and fewer firm answers than almost any other.

In outline, the process is this: A complete RNA transcript of the coding and noncoding mosaic is made by RNA polymerase. A small chemical residue, or a cap, is added to the front, 5′ end, of the transcript. A string of around 200 adenine molecules, the poly(A) tail, is added to the back, 3′ end, of the transcript. The noncoding regions are cut out. The coding regions are stitched back together in the same order as they appeared in the DNA. And the mature message exits into the cytoplasm where protein synthesis may begin.

Questions relating to RNA processing may be divided into two areas: those relating to the excision of the intervening sequences; and those concerning the interaction of the RNA transcript with other molecules, such as proteins.

(Continued on page 32)

Littlewood Conjecture Resolved

In 1948, British mathematician John E. Littlewood published a surprising conjecture. No matter how clever you are at superimposing sine and cosine waves of different frequencies and amplitudes greater than or equal to 1, he said, you can never get them to completely cancel each other out. In fact, when you add these waves together and integrate, the area they bound approaches infinity as the number of waves being added gets larger and larger.

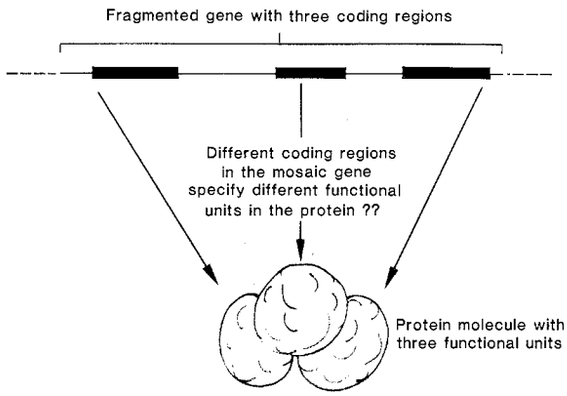
Ralph Boas, Northwestern University

John E. Littlewood 1938

Littlewood's conjecture has intrigued mathematicians for the past 33 years, in part because it seemed so difficult to verify or disprove and in part because it has applications in harmonic and functional analysis. Progress was made in 1959 by Paul Cohen of Stanford University, in 1979 by John Fournier of the University of British Columbia, and, last year, by S. K. Pichorides, who is visiting at the University of California at Los Angeles. Both Cohen and Pichorides were awarded prizes for their work. But no one was able to show, as Littlewood conjectured, that the integral of the absolute value of the sum of N sine and cosine functions must be at least a constant times the logarithm of N.

Recently, Brent Smith of Illinois State University, Louis Pigno of Kansas State, and O. Carruth McGehee of Louisiana State succeeded in verifying the conjecture. Their paper describing their result has been accepted for publication in the *Annals of Mathematics*. As the three mathematicians are quick to point out, their proof

0036-8075/81/0403-0028$00.75/0 Copyright © 1981 AAAS



Rationale for fragmented genes

One attractive explanation of the piecemeal arrangement of genes from nucleated cells is that the coding regions represent discrete functional regions in the protein for which the entire gene codes. Such an organization would have important evolutionary implications.

(Continued from page 30)

Two things can be said about the excision of intervening sequences: the mechanism by which it is done is extremely precise; no one has any good idea how it is achieved.

Close scrutiny of a large number of intervening sequences should, it might reasonably be expected, give some insight into mechanisms. But it does not. Intervening sequences range in size from 10 to 10,000 bases long, and there is no apparent pattern in how they are spaced in the mosaic gene. All that can be said is that at the beginning of most intervening regions is a nucleotide doublet GT, and at the end another doublet, AG. These doublets mark the splice junctions. They are clearly important. But by themselves they cannot account for the precision required in, say, removing 20 intervening sequences from a transcript so that the coding regions are joined together in the correct order, and so that the message encoded in the sequence of bases is not perturbed in any way.

"This is a real enigma," admits Dean Hamer of the National Cancer Institute. The enigma becomes more profound with the following kinds of experimental results. Remove the intervening sequences from some genes and they will fail to be transcribed or are transcribed at low levels. Other genes do not require intervening sequences in order to be transcribed. Some RNA transcripts that have experimentally removed intervening sequences fail to be transported normally to the cytoplasm. It is possible to delete all but about 15 or so nucleotides at both splice junctions from an intervening sequence, and it is still excised accurately and the adjacent coding regions are spliced. It is possible to create a hybrid intervening sequence—from, for instance, a virus and a mouse globin gene—and excision and splicing go on quite normally. "There are many paradoxes," says Paul Berg of Stanford.

The secrets to splicing may lie in the structures with which the RNA transcript is associated in the nucleus. What are these structures? "There's a cytoskeleton, a kind of proteinaceous matrix, within the nucleus," says O'Malley, "and you often find the RNA transcripts associated with this." Additionally, or perhaps as part of the cytoskeleton, there are large conglomerations of small pieces of RNA and a bundle of core proteins: these are the ribonucleoprotein particles. Conceding that the field is more than a little controversial, Terence Martin, of the University of Chicago, explains that the particles are made up of around 25 protein molecules (combinations of six different proteins), that the bead they form is about 250 angstroms in diameter (accommodating about 1000 RNA bases), and that electron microscopy and other techniques reveal that newly transcribed RNA is usually packaged with the ribonucleoprotein complex. "This is the real substrate for RNA processing," says Martin.

There is an obvious and seductive parallel between the string-of-beads structure of DNA and protein cores in chromatin and possible structural arrangements between RNA transcripts and the ribonucleoprotein particles. "But," insists Roger Kornberg, of Stanford, "I don't find compelling evidence for this as a substrate structure." Opinions here differ widely.

> "Recombinant DNA technology has allowed us to accumulate a massive amount of data . . ."

Whatever is the structure of the processing substrate, it is difficult to see how regular wrapping of an RNA transcript could be a direct aid in the excision of intervening sequences that show no regularity whatsoever. The ribonucleoprotein particle, or the nuclear matrix, might merely provide a bench on which other agents can act. About 8 years ago a group of small RNA molecules was discovered in cell nuclei; since then they have been without an ascribed function. A year ago Joan Steitz, of Yale, discovered that the sequence of some of these small RNA molecules is somewhat complementary to the sequences at the splice junctions of some intervening sequences. She suggested that these molecules might act as guides to bring the splice junctions together, allowing precise excision and rapid subsequent splicing.

"Many people find this suggestion attractive," says O'Malley. "It's the best thing we have to go on at present," suggests James Darnell, of Rockefeller University. "Too simplistic," comments Berg. Even if these small nuclear RNA's are involved in splicing, it is still difficult to see how they could ensure that excision mistakes do not occur between adjacent intervening regions.

If the mechanism of splicing is unclear, the question of what is involved in transporting the spliced message from the nucleus to the cytoplasm is a total mystery. Some new way of tackling the problem is desperately needed.

The large number of steps involved in gene expression—from initiation of transcription to translation of the message into protein in the cytoplasm—gives an insight into the number of levels at which control may be exerted over gene function. "People had taken it for granted that control of expression might take place at many levels," Darnell told *Science*, "but what we saw at the San Francisco meeting showed that multilevel control is a reality. It's clear too," he continued, "that control at the level of transcription is the dominant system."

An uninvolved outsider cannot fail to be impressed with the precision and confidence with which biologists can now manipulate genetic material. "Recombinant DNA technology has allowed us to accumulate a massive amount of data in a relatively short time," says Baxter, who helped organize and was official chairman of the San Francisco conference. "Understanding the way genes are regulated is the biggest challenge since the discovery of the double helix, and you can see we are making great progress."—ROGER LEWIN