Gene Control Puzzle Begins to Yield

At long last investigators are getting their first look at the DNA segments that help to turn eukaryotic genes on and off

The discovery of restriction enzymes and the development of recombinant DNA techniques have made possible gene manipulations that were barely dreamed of in the not too distant past. "Today it is easier to work with DNA than with any other macromolecule, which is the exact reverse of the situation 10 years ago," says Donald Brown of the Carnegie Institution of Washington in Baltimore.

The latest application of the techniques—and one that is sweeping the molecular biology community, if presentations at a recent symposium* are any guide—is for the identification of the DNA segments that help control the turning on and off of gene expression in the nucleated cells of higher organisms (eukaryotes).

Although an understanding of eukaryotic gene control is needed to gain insight into the development of specialized tissues in higher organisms and into such problems as cancer, which may involve inappropriate gene expression, it has lagged far behind knowledge of the much simpler nonnucleated cells of bacteria. The new work has not yet answered all the outstanding questions—practically nothing was known about eukaryotic gene control to begin with—but it has provided the first clues. And there are already indications that gene regulation in nucleated cells may differ significantly from that in bacteria, even though there are some resemblances between the two.

One of the early surprises came from Brown's laboratory. For the past 10 years he has been studying the structure and expression of a small eukaryotic gene that codes for one of the RNA components of ribosomes, the intracellular factories where protein synthesis takes place. Both the gene and the RNA copied from it, which is called 5*S* RNA because of the way it sediments during ultracentrifugation, consist of 120 nucleotides.

The surprise, which Brown reported early in 1980, was that the control region for turning on the 5*S* RNA gene is right in the middle of the gene. This was the first example of such an arrangement. All the known control regions for bacterial genes had been found outside the gene proper, in the DNA segments lying just before the initiation site. This is the nucleotide on the 5′ end of the gene where transcription into RNA copies begins.

The methods used by Brown and his colleagues to locate the 5*S* gene control site exemplify those generally used to spot gene control regions. "After you isolate the gene," Brown explains, "you progressively take off pieces and clone copies of each shortened form. Then you test the altered genes to see if they make the right gene product. . . . In this way, you can do in vitro genetics with a piece of DNA." This procedure would not have been possible without the development of recombinant DNA techniques and of systems for assessing the activity of the altered genes, which also became available only in the past few years. The assays can be done in vitro, by adding the genes to cell-free extracts that can accurately transcribe them, or in vivo, by introducing genes into living cells, where they are also expressed (*Science*, 19 December 1980, p. 1334).

> . . . there are already indications that gene regulation in nucleated cells may differ significantly from that in bacteria . . .

The Carnegie workers isolated the 5*S* RNA gene from frog cells, where it is present in multiple copies separated by short sequences of spacer DNA. According to Brown, "We thought that when we took off all the spacer DNA we would kill the gene, but we didn't." Instead, they found that the gene was not inactivated until they cut off the first 55 nucleotides, working in from the 5′ end. When they began cutting from the other, 3′ end, the gene did not become inactivated until they had removed about 40 nucleotides. This means that the nucleotides from about position 55 to position 80 in the gene are the ones needed for initiation of transcription—not the nucleotides on the spacer sequences flanking the 5′ end, as might have been predicted from the bacterial results. "But the flanking region is not irrelevant," Brown says. "It has a subtle effect on transcription, influencing the exact start site of transcription." When the region is deleted, transcription may be initiated one to a few nucleotides away from the normal starting point.

The turning on of gene expression depends on an interplay between the gene itself and other cellular components. One of these is the enzyme that actually transcribes the gene, producing an RNA product. In the case of the 5*S* RNA gene, this enzyme is RNA polymerase III (pol III), which has a limited range of action. The few genes it has been found to transcribe include, in addition to the one for 5*S* RNA, the genes for the transfer RNA's, one of the adenovirus genes, and some repeated eukaryotic gene families of unknown function.

How the control sequence identified by Brown directs pol III to begin making 5*S* RNA is unclear, but it is known that additional cellular factors are required. While Brown has been cutting up genes to see which sequences are important for their control, Robert Roeder of the Washington University Medical School has been taking a totally different approach to the problem. "Most of what we have done has not involved sequence modification," he points out. "Basically, we have been dissecting the cell-free systems to identify protein factors that mediate transcription." Gene control se-

*The symposium, entitled Developmental Biology Using Purified Genes, was sponsored by ICN Pharmaceuticals and the University of California at Los Angeles. It was held in Keystone, Colorado, on 15 to 20 March.

0036-8075/81/0508-0653$00.75/0

quences might then be identified by seeing which DNA segments interact with the factors.

Roeder has so far identified three cellular protein factors that are needed for the activity of pol III. One of these, which has been purified and shown to have a molecular weight of 40,000, has turned out to be specifically required for initiation of transcription of the 5*S* RNA gene. "This protein," Roeder says, "binds to the central region of the gene. Our binding site overlapped exactly with the control region identified by Brown." Both Roeder and Brown have shown that in the absence of this factor, transcription of the 5*S* RNA gene cannot begin. The factor is not needed for transcription of the other genes copied by pol III.

The remaining two protein factors, which have not yet been purified and characterized, are needed for transcription of all the pol III genes. Roeder concludes, "There are multiple factors; some are common and some may be gene specific." Specific factors for transcription of pol III genes, other than the 5*S* RNA gene, have not been identified, if there are any. Roeder says the possibility still exists that they might be lurking in the two unpurified protein factors, however.

If the range of genes transcribed by pol III is limited, that of genes transcribed by RNA polymerase II (pol II) is very wide, as it includes all the genes that code for proteins. "Pol II is *the* enzyme for transcribing unique genes," says Philip Leder of the National Institute of Child Health and Human Development. These include the genes that are expressed in specific, differentiated cells, such as muscle or nerve cells, and that must be turned on or off at appropriate times during development. Not surprisingly, a large number of investigators are currently engaged in trying to identify the control regions for the pol II genes.

About 2 years ago, Michael Goldberg, who was then a graduate student in the laboratory of David Hogness at Stanford University, obtained a clue to pol II control while studying the structure of the histone genes of the fruit fly. He identified a sequence of ten or so nucleotides, mostly containing the bases thymine (T) and adenine (A), that began about 30 nucleotides to the left of the 5′ end of the genes. This sequence is called the TATA box, because of its nucleotide composition, or the Hogness or Goldberg-Hogness box. So far, TATA boxes have been identified in similar locations in nearly 60 eukaryotic genes from a variety of species, which are transcribed by pol II. Conservation of a DNA sequence in this way usually means that it has some essential function.

The nucleotide sequence of the TATA box closely resembles that of a gene segment known to be part of the bacterial promoter region. Promoters are usually defined as DNA sequences that are indispensable for the expression of structural genes. The fact that the TATA box and the bacterial gene segment have similar structures and occupy analogous positions in their respective genomes suggested that the TATA box might be a eukaryotic promoter.

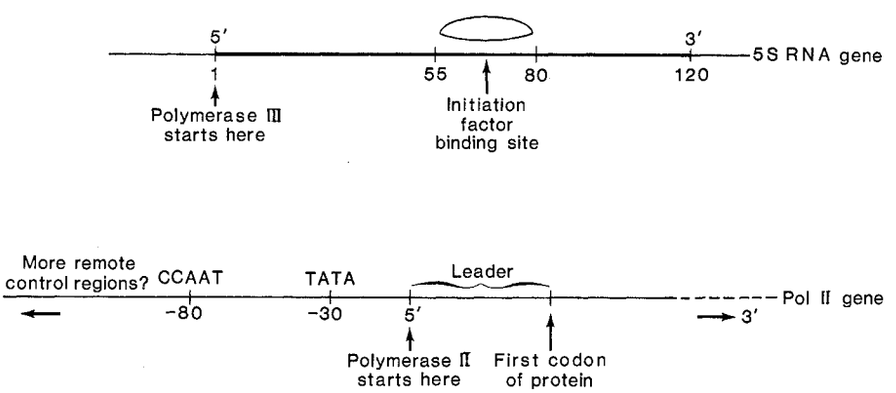
And this appeared to be the case—for a while. Several investigators showed that deletions of the TATA boxes of a number of different genes prevented, or at least drastically reduced, transcription in vitro. Such deletions do not just remove the TATA box, however; they also bring upstream nucleotides much closer to the beginning of the gene than they would otherwise be. Consequently, the decreased transcription might be caused not by loss of the TATA box, but by addition of some inhibitory sequence.

Pierre Chambon of the University of Strasbourg effectively ruled out this possibility in an experiment in which he introduced a single mutation into the TATA box of the gene for the protein conalbumin. After TATA was converted to TAGA by the substitution of a guanine (G)-containing nucleotide for the second thymine-containing nucleotide, in vitro transcription of the gene was almost abolished.

Nevertheless, the simple picture of the TATA box as a eukaryotic promoter, which seemed to be favored by the in vitro experiments, became more complicated as results began to come in from studies in which the altered DNA's were put into living cells to be transcribed. For example, Steven McKnight, who is also at the Baltimore branch of the Carnegie Institution of Washington, found that the TATA box was not required for in vivo transcription of the *tk* gene—a viral gene for the enzyme thymidine kinase. When he excised it from the gene and introduced the altered DNA into frog oocytes or cultured mouse cells, accurate transcription was not significantly decreased. Other investigators, notably Max Birnstiel of the University of Zurich and Chambon, also have evidence suggesting that the TATA box is not essential for initiation of transcription in living cells.

This does not mean that it has no role, however. Their work shows that this DNA segment is needed for selection of the correct starting sites, a finding similar to that for the 5′ flanking region of the 5S RNA gene. As Chambon describes the situation, "The role of the TATA box is to direct the RNA polymerase to initiate about 30 bp [base pairs] downstream. In the absence of the TATA box initiation occurs at multiple sites. . . ." Under normal circumstances, transcription begins, accurately, at only one or a few specific sites.

Research from McKnight's laboratory



Two kinds of genes and their proposed control sites

The upper diagram shows the 5S RNA gene of the frog, which is transcribed by RNA polymerase III. Two control sites cooperate to produce accurate initiation of the transcription of this gene. The interior site, after binding a cellular initiation factor, somehow directs the turning on of the gene, while the DNA segment just to the left of the 5′ initiation site helps to ensure an accurate start. The pol II gene, shown in the lower diagram, also requires the cooperation of at least two gene segments for transcription to begin in living cells. The TATA box performs a function similar to that of the 5′ flanking region of the pol III gene. More remote regions, including the CCAAT box, help to determine the efficiency of transcription. Transcription of practically all pol II genes begins with a "leader" sequence that codes for amino acids not found in the final protein product. Not shown is the fact that most of these genes are interrupted by one or more—and often many—noncoding DNA segments.

and those of several other investigators shows that DNA sequences some distance upstream from the TATA box may be needed for the in vivo expression of pol II genes. Perhaps the safest thing to be said about the research at this time is that the results have not all been sorted out.

Some results that need such sorting out are those concerning the CCAAT box (C represents nucleotides containing the base cytosine), a conserved DNA sequence found about 80 nucleotides upstream from the initiation sites of many genes, and another candidate for a regulatory site.

In McKnight's studies of the *tk* gene, he found that a region consisting of about 60 nucleotides centering around the CCAAT box are necessary for efficient initiation of transcription of the gene. Deletions of DNA segments into that region effectively abolished accurate transcription. In contrast, Birnstiel found that deletion of the CCAAT box increased the transcription of a histone gene about 20-fold.

DNA segments even farther upstream than the CCAAT box have also been implicated in gene control by several in vivo studies. The Birnstiel group found that deletion of a sequence beginning about 120 nucleotides upstream from the histone gene initiation site greatly reduced transcription.

Another example comes from studies of a viral genome, that of SV40, which is frequently used as a stand-in for the cellular genes because it is transcribed by the same enzyme in infected cells. This virus has a 72-nucleotide segment that occurs as a tandem repeat. The double unit extends from position 116 to position 261, counting upstream from the initiation site for the SV40 early genes. (The SV40 genome is transcribed in two sections, one activated shortly after infection and the other several hours later.) Chambon and also George Khoury and Peter Gruss of the National Cancer Institute have shown that at least one copy of the repeated unit is required for transcription of the SV40 genome in infected cells. Chambon concludes, "Sequences located upstream from the SV40 early gene TATA box are crucial for efficient in vivo initiation of transcription."

Moreover, Robert Tjian of the University of California at Berkeley, by systematically deleting segments of the SV40 genome, has shown that a segment that includes a portion of one of the repeats is needed for transcription in vitro of the early genes. This is one of the few examples of such a requirement. In most other studies, transcription in vitro has not appeared to require participation of the remote sequences, but only of the TATA box.

The evidence obtained thus far suggests that for the SV40 early genes, and perhaps for many other genes transcribed by pol II, the TATA box helps to determine the exact starting site while the more remote DNA segments help to determine the efficiency of transcription. This picture resembles that for pol III transcription. Although the 5*S* RNA gene does not have a TATA box, the 5′ flanking region performs a similar function. The main difference is that for this gene the remote control site is in the interior whereas for the pol II genes it lies outside.

There are several hypotheses to explain how the remote sequences might affect initiation, but the favorite involves participation of the chromatin structure of DNA. The DNA found in the cell is not naked, but comes clothed in proteins, mainly histones. The complex of DNA plus protein produces a structure resembling a string of beads, with the protein forming the beads and the DNA coiling around and extending between the beads. There are indications that the DNA around the beads is less accessible to enzymes, presumably including polymerases, than the bare DNA of the string. So one explanation for the effect of remote DNA segments on transcription efficiency is that they position the beads in such a way that the initiation site is exposed to polymerase. Alternatively, the coiling of the DNA might bring together separate DNA regions that must interact to bind the polymerase, or foster some other aspect of initiation.

Involvement of these higher structural features may help to account for the difference between the in vitro and in vivo results. Most investigators consider that chromatin formation by DNA introduced into cell-free extracts is unlikely. Roeder, who was one of the pioneers in developing the cell-free systems, says, "I don't think there is any major discrepancy between the two kinds of results. Transcription is a multistep process. We may be looking at only a part in cell-free systems." In other words, in studying what happens to naked DNA, investigators may be bypassing some of the earlier events needed to initiate transcription.

. . . the TATA box helps to determine the exact starting site while the more remote DNA segments help to determine the efficiency of transcription.

Even results with genes introduced into living cells probably do not fully reflect the normal sequence of steps by which genes are activated. Most of these genes have been introduced into cells where they are not normally expressed—globin genes put into fibroblasts is an example—but the transferred genes are transcribed nonetheless.

Investigators are devising systems in which transferred genes are activated in response to the normal inducers, a situation that may more accurately reflect what happens in nature. For example, Richard Axel of the College of Physicians and Surgeons of Columbia University has introduced a heat-shock gene of the fruit fly into mouse cells. The gene appears to be transcribed into RNA when—and only when—the cells are exposed to high temperatures, just as it is in the fruit fly donor. In addition, Charles Weissmann of the University of Zurich has shown that interferon genes that have been introduced into cultured cells are turned on in response to virus infection, just as they normally are. By systematically modifying genes such as these before they are transferred, investigators should be able to make further progress in identifying gene control sites and understanding how genes are turned on or off in nucleated cells.

—JEAN L. MARX