A Crystalline View of Protein-DNA Binding

Studies of the first co-crystals of DNA-binding proteins and their target DNA's help to reveal how they interact with one another

By binding to specific DNA sequences, regulatory proteins help to turn genes on or off as required. These interactions are crucial for maintaining the everyday life of all cells and for guiding the development of complex, multicellular organisms. Not surprisingly then, molecular biologists are very interested in understanding how proteins are able to pick out particular, short DNA sequences from an entire genome.

Over the past few years, evidence from several laboratories has indicated that there is a common pattern by which the regulatory proteins of bacteria and the viruses that infect bacteria (bacteriophages) recognize their target sequences. Now, John Anderson, Mark Ptashne, and Stephen Harrison of Harvard University have produced co-crystals of the DNA-binding region of a bacteriophage regulatory protein and the DNA sequence it recognizes. Using x-ray crystallographic methods, the Harvard workers have determined the three-dimensional structure of the complex to a resolution of 7 Å (*1*). The results provide direct confirmation of the model that grew out of the earlier work.

Although this is the first report of the structure of a co-crystal of a regulatory protein and its target sequence, John Rosenberg of the University of Pittsburgh School of Medicine and his colleagues had previously solved to a 3-Å resolution the structure of a co-crystal of Eco RI, a bacterial restriction enzyme, with its target DNA sequence (*2*). Restriction enzymes such as Eco RI bind to specific DNA segments, as the regulatory proteins do, but the enzymes cut the DNA instead of activating or repressing gene expression. Analysis of the Eco RI co-crystal shows that the binding region of the enzyme resembles those of the regulatory proteins to a degree but also displays significant differences.

Early indications that bacterial regulatory proteins could have similar recognition sites for DNA came in 1981 and 1982 when investigators first determined the three-dimensional structures of a number of the uncomplexed proteins. Wayne Anderson of the University of Alberta in Edmonton, Brian Matthews of the University of Oregon, and their colleagues solved the structure of the Cro protein, which is produced by bacteriophage λ and is a gene repressor. The structure of the catabolite gene activator protein from the bacterium *Escherichia coli* was solved by Thomas Steitz and his colleagues at Yale University. And Carl Pabo and Mitchell Lewis, who were then working in Harrison's and Ptashne's laboratories, obtained the structure of the DNA-binding region of another λ regulatory protein, the λ repressor which shuts off all of the viral genes except one. It activates the expression of its own gene.

By binding to their target regulatory sequences, which are called operators, these proteins either activate or prevent the transcription of the associated genes into messenger RNA. The operator sequences are symmetrical, as are the structures of the proteins themselves. They are all dimers, consisting of two identical protein chains.

The interactions of the bacterial and phage regulatory proteins with DNA follow a general pattern.

Although the overall amino acid sequences of the individual chains bear little resemblance to one another, the crystallographic analyses revealed a striking similarity in their three-dimensional structures—one that identified the probable location of the DNA recognition site. Each protein contained a "two-helix motif," consisting of two α-helices connected by a short nonhelical segment. The angle formed by the two helices was the same and the structures were virtually superimposable for all three proteins.

One of the α-helices projects out from the surface of the molecule. The complete protein, with its two protein subunits, would thus have two such projecting helices.

Comparison of the protein structures with that of double-helical DNA in the right-handed B-DNA conformation, the "ordinary" Watson-Crick structure, showed that the projecting α-helices could fit neatly into two successive sites along the major groove of the B-DNA double helix where the amino acids on the outer surfaces of the helices can make contact with the bases of the recognition site. The putative recognition helices are too close together to fit into successive grooves of the Z-DNA conformation, a structural form that was originally identified by Alexander Rich and his colleagues at the Massachusetts Institute of Technology. The second helix of the two-helix motif lies across the major groove of the B-DNA.

Only by determining the three-dimensional structures of the protein-DNA complexes themselves could investigators show the model for the interaction to be correct. That is what Anderson, Ptashne, and Harrison have now done by their x-ray studies of a co-crystal of the DNA-binding region of the repressor protein of coliphage 434 and the 14-base pair DNA sequence that the protein recognizes. Earlier work had indicated that the 434 and λ repressors are similar, although the 434 protein lacks an armlike segment that the λ repressor wraps around the DNA. In any event, the 434 repressor interacts with DNA just as the model predicted and its binding produces very little deformation of B-DNA.

Moreover, Robin Wharton, also of Harvard, and Ptashne have recently completed a feat of protein engineering, the results of which provide additional support for the proposed model and indicate that it may be possible to identify the DNA recognition sites of proteins from their amino acid sequences (*3*). Examination of the sequences in the two-helix motifs had already shown that some of the amino acids, in particular those that are needed to maintain the three-dimensional structure, are conserved among the different proteins even though their amino acid sequences are otherwise dissimilar. In contrast, the outer, solvent-exposed amino acids of the recognition helix of the regulatory proteins, the ones that make the specific contacts with the target DNA, vary from one protein to another. Wharton and Ptashne replaced the amino acids of the 434 repressor that were predicted to be involved in DNA recognition with the corresponding amino acids of another repressor protein, this one from a bacteriophage that infects *Salmonella*. They left intact the conserved amino acids that maintain the two-helix structure.

The Harvard workers reasoned that the engineered 434 repressor protein would still fold normally, but that its specificity would be changed. That is what they found. The altered protein bound to the operator of the *Salmonella* phage DNA just as avidly as the *Salmonella* repressor protein, but it no longer recognized the 434 operator.

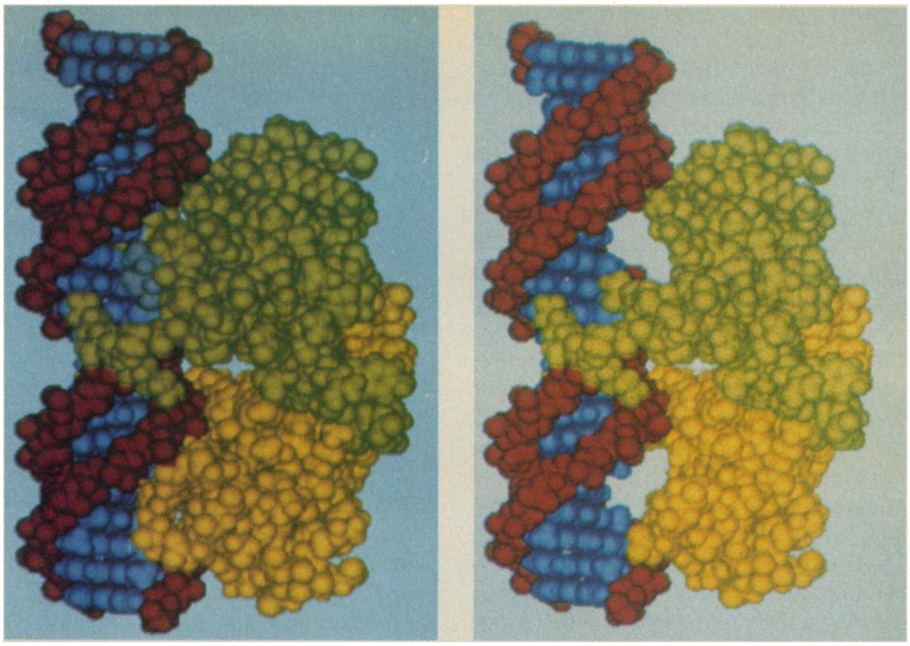
All in all, the data from the various laboratories provide strong evidence for the hypothesis that the interactions of the bacterial and phage regulatory with DNA proteins follow a general pattern. The model may soon be extended to include the *trp* repressor of *E. coli*, according to Paul Sigler of the University of Chicago.

The Rosenberg group finds that binding of the Eco RI restriction enzyme to its target DNA sequence is in some ways similar to the binding of the regulatory proteins. For example, the enzyme is also a symmetrical dimer that recognizes a symmetrical segment of DNA, although this one contains only 6 base pairs compared to approximately 15 in the operators. In addition, the Eco RI binding region has an α-helix that fits into the DNA major groove. However, the binding region has just one helix and the angle that the helix makes with the DNA molecule is different from that made by the recognition helices of the regulatory proteins.

But the most significant difference between Eco RI binding to DNA and that of the regulatory proteins is the distortion induced in the DNA structure by the restriction enzyme. When Eco RI binds to its target site, it causes the double-stranded molecule to rotate about 25°, as a result of which the strand backbones move approximately 4 Å farther apart. This kink, as Rosenberg calls it, helps Eco RI solve a problem.

The DNA strands must be forced apart to accommodate the protein's recognition helix, but directly breaking the hydrogen bonds that hold the bases of the two DNA strands together would require a great deal of energy. By causing a partial unwinding of the DNA double helix, the restriction enzyme separates the two DNA chains in an energetically more favorable way.

According to Rosenberg, discovery of the kinked form of DNA in the complex with Eco RI extends the repertoire of known DNA conformations, which currently include the A and C forms in addition to the B and Z forms. Rich had previously suggested that protein binding can stabilize DNA in the Z conformation. "There may be a class of DNA conformations that only appear when proteins bind," Rosenberg says.



Interaction of B-DNA with a dimer of the DNA-binding regions of the λ repressor. In the view on the right, most of the two-helix motif has been deleted. [M. Ptashne, Trends Biochem. Sci. *9, 142 (1984)]*

As yet comparatively little is known about the proteins that regulate gene expression in higher, eukaryotic organisms. Although investigators have in recent years identified a number of proteins that may participate in eukaryotic gene control, most of these have not been fully characterized and it is too early to say whether the eukaryotic proteins follow the same recognition patterns as the prokaryotic regulators.

One eukaryotic regulatory protein that has been well characterized is transcription factor IIIa (TFIIIa), which has been studied extensively by Robert Roeder of Rockefeller University and his colleagues and by Donald Brown's group at the Baltimore branch of the Carnegie Institution of Washington. This protein is needed for gene transcription by polymerase III, which acts on a limited number of genes, including those coding for one of the ribosomal RNA's.

TFIIIa is very different from the prokaryotic regulatory proteins. It consists of one protein chain with nearly 350 amino acids, 300 of which form an elongated DNA-binding region. Recent analysis of the TFIIIa structure by Aaron Klug's group at the Medical Research Council Laboratory of Molecular Biology in Cambridge, England, points up the unusual structure of the binding region (*4*). It apparently consists of nine globular domains, each containing zinc and separated by short flexible DNA regions. Brown describes the structure as something like a caterpillar with nine segments.

TFIII binds to the ribosomal RNA genes more or less permanently. It allows repeated rounds of transcription, even though it attaches in the middle of the genes where it might block the passage of the polymerase. Klug and his colleagues suggest that this does not happen because individual segments of the protein separate from the DNA as the polymerase moves along, while other segments maintain their hold.

Most eukaryotic genes, including those that code for proteins, are transcribed by polymerase II. Although the evidence is still largely indirect, some of the proteins that regulate the expression of these genes bear at least a superficial resemblance to the prokaryotic regulators. This is true, for example, of the large T antigen of SV 40, which represses the transcription of viral genes.

In addition, Carl Parker and his colleagues at the California Institute of Technology have identified a protein needed for transcription of a heat shock gene in the fruit fly *Drosophila melanogaster* (*5*). (Heat shock genes are activated by heat and other stresses and are found in species ranging from bacteria to man.) The *Drosophila* protein is a dimer, like the prokaryotic proteins, and the DNA sequence to which it binds displays the same kind of symmetry as the recognition sequences in the lower organisms.

Moreover, aspects of the binding of the regulatory protein of the heat shock gene resemble λ repressor binding. In both cases, the regulatory regions of the genes contain multiple copies of the target DNA sequence. The proteins must bind to at least two of these to produce their effects and binding to the first site facilitates binding to the second.

Robert Tjian and his colleagues at the University of California at Berkeley have identified a protein that binds to one of the regulatory regions of the SV 40 genome and is needed for transcription of the associated viral genes (*6*). Several cellular genes also respond to the protein, which is called Sp1. Indirect evidence suggests that the mode of contact of Sp1 and its target sequence on DNA may resemble the interaction between the prokaryotic factors and their target sequences.

There are major differences, however. Sp1 appears to contact only one DNA strand, not two, and the sequence recognized by the protein is not symmetrical. The lack of symmetry raises an interesting paradox, Tjian notes. Binding of Sp1 activates transcription no matter what the orientation of the target DNA segment and it is difficult to imagine how an asymmetric sequence might work in both directions. The results imply that Sp1 might itself be symmetrical.

Control of SV 40 gene expression involves more than one type of regulatory sequence. Sp1 acts on the promoter, which is needed for accurate initiation of transcription. Other proteins that bind to the SV 40 enhancer are being identified. Direct participation of these proteins in transcription has yet to be conclusively demonstrated, but if some or all are involved then gene control in eukaryotes would be significantly more complicated than in prokaryotes.

Even more intriguing are indications that the peptide encoded by the "homeo box," a DNA segment that has been linked to a number of developmentally important genes of the fruit fly and is also found in mammalian genomes, has an amino acid sequence that may have the capacity to fold into the two-helix motif. Fruit-fly proteins that bear the homeo-box sequence have been implicated in gene control during development. A DNA binding site would be consistent with this hypothesis.

In view of the complex life cycles of higher organisms, it would not be especially surprising if eukaryotic gene control is more complex than that in prokaryotes, as the SV 40 work implies. Bacteria after all do not have the developmental problems of the eukaryotes, which must turn specific genes on or off in particular cell types as they grow from single cells to multicellular organisms. That problem has long intrigued investigators, who are now finding new clues in the proteins that interact with DNA.

—JEAN L. MARX

References and Notes

1. J. E. Anderson, M. Ptashne, S. C. Harrison, *Nature (London)* **316**, 596 (1985).
2. C. A. Frederick *et al.*, *ibid.* **309**, 327 (1984).
3. R. P. Wharton and M. Ptashne, *ibid.* **316**, 601 (1985).
4. J. Miller, A. D. McLachlin, A. Klug, *EMBO J.* **4**, 1609 (1985).
5. C. S. Parker and J. Topol, *Cell* **37**, 273 (1985).
6. D. Gidoni, W. S. Dynan, R. Tjian, *Nature (London)* **312**, 409 (1984).

Tracking a Stormy Beast in the Night

Weather satellites have revealed thunderstorms organized into unexpectedly large nighttime rainstorms over the central United States

The thunderstorms were ordinary enough to begin with, if rather severe. They broke out during midafternoon of 23 June in a line from south-central Iowa across southeastern Nebraska and into northeastern Kansas. For several hours the storms, especially those over Lincoln, Nebraska, unleashed tornadoes, wind gusts to over 130 kilometers per hour, hail as big as baseballs, and up to 10 centimeters of rain.

That kind of weather is not unusual on a summer afternoon in the central United States, but these storms did not stop with a bit of brief, locally severe weather. By 9 p.m. that evening, they had somehow created a single, roughly circular rainstorm 200,000 square kilometers in area. Outlasting smaller storms from that afternoon, it persisted until dawn, when only a lingering swirl of clouds remained. But then, again unlike an everyday thunderstorm, its remnants lingered into the next afternoon, when the heat of the day rejuvenated it so that once again it unleashed severe weather, this time over southern Illinois and western Kentucky.

Until a few years ago, meteorologists would have likely attributed this weather to a number of big but scattered thunderstorms. Now the perspective gained through weather satellites has shown that by somehow organizing the circulation of the atmosphere around them, some thunderstorms can transform themselves into larger, longer-lived rain systems, perhaps through processes found in tropical storms. This newly recognized organization of thunderstorms is called a mesoscale convective complex, or MCC. However the transformation occurs, the size and longevity of MCC's offer meteorologists a better chance to forecast the sometimes disastrous, sometimes life-giving, rains of the summertime central United States.

What sets an MCC apart from other kinds of summer thunderstorms, at least from the point of view of those getting rained on, is how long the rain continues. The rain from storms along a rapidly moving front is heavy but brief as the front passes by quickly. It is the front's motion, in fact, that plows warmer, moist air upward ahead of it to produce rain. Once started upward, the air expands and thus cools, condensing some of the moisture. That condensation leads to rain as well as still more vertical motion through the heat it releases—warmer air is lighter, more buoyant. Similar vertical motion or convection can also pop up here and there on a warm, humid summer afternoon due simply to random atmospheric disturbances. In any case, the rainfall of the resulting scattered storms is also limited to a small area and is usually brief.

The rain from MCC's, on the other hand, can be heavy at times, cover a large area, and go on for hours. These storms get their start as typical convective systems during warm summer afternoons east of the Rockies. But then, unlike systems in the West or East, they grow into a different sort of storm. A single system of convection may expand or a group of scattered storms may merge.

The new storm will have one, two, or even three lines of towering convective cells embedded in it that can produce all the damaging weather for which severe thunderstorms are renowned. Ray McAnelly of Colorado State University has found that 12 MCC's in his study