DNA recognition, warts and all

John Kuriyan and Stephen K. Burley

THROUGHOUT the life of a cell, its genetic blueprint is consulted and interpreted in response to various developmental and environmental signals. Central to this process is the recognition of specific DNA sequences by transcription factors, which then determine whether or not a particular gene is used. What are the mechanisms utilized by these DNA-binding proteins to enhance sequence-specific interactions? The answer is beginning to emerge from the analysis of the structures of protein–DNA complexes determined by X-ray crystallography and nuclear magnetic resonance.

On page 505 of this issue[1] Hegde *et al.* present the three-dimensional structure of the complex between a papillomavirus transcription factor and DNA, which reveals a new protein architecture as well as a new template for DNA–protein interactions. This DNA-binding fold and other previously determined structures (for reviews see refs 2 and 3) show that although the DNA comprising the genetic library is a conservative edifice, rich in information content but poor in structural variation, the proteins that serve as librarians are not predictably dull. Rather, transcription factors display all the architectural flamboyance characteristic of globular proteins.

Regions

The papillomaviruses are a family of DNA viruses that cause warts in humans and other mammals (for review see ref. 4). Proteins encoded by the viral gene E2 play central parts in the regulation of transcription and viral DNA replication, and are thought to consist of three functionally independent regions. The amino- and carboxy-terminal segments are highly conserved and constitute the activator and DNA-binding segments, respectively, while a central, non-conserved region appears to form a spacer between them. The 85-residue DNA-binding region will bind to specific DNA sites with high affinity, and is the subject of the X-ray structural work of Hegde *et al.*[1].

Like many viral transcription factors, E2 protein binds to palindromic DNA sequences as a dimer, presumably as a means of maximizing specificity while minimizing the size of the E2 gene. It does not, however, resemble any known DNA-binding module, such as the helix–turn–helix, zinc-containing, leucine zipper or β-ribbon proteins[2,3,5]. Rather it forms a dimeric eight-stranded antiparallel β-barrel structure (see Fig. 2, page 507) that presents, on the surface of the barrel, two symmetrically disposed α-helices (the DNA-recognition helices). Although β-barrels are a common architectural feature in proteins (witness the number of enzymes that contain the β-barrel found in triose phosphate isomerase), E2 is the first example of a structure in which a barrel is formed by the dimerization of two polypeptide chains. This feature explains why the dimer is so exceptionally stable and cannot be dissociated, except under denaturing conditions. Each polypeptide chain contributes half of the hydrophobic core that forms the interior of the dimeric barrel. Two monomers therefore have a strong incentive to pair and exclude water, and the dimer is further stabilized by hydrogen bonds that form the seams of the barrel.

The two recognition helices are positioned such that both cannot make simultaneous contact with linear B-form DNA. Instead, the DNA bends more or less smoothly around the E2 protein into an arc that engages both α-helices within the major grooves (for comparison, this curvature is comparable to that of DNA spooled about the nucleosome core particle in chromatin). Deformations from ideal B-DNA character are now emerging as a common feature of protein–DNA interactions, but it remains to be seen whether sequence-dependent variation in the deformability of DNA is a general determinant of the specificity of DNA–protein interactions.

E2 interacts with 17 highly conserved target DNA sequences in the papillomavirus genome, and the crystal structure shows a network of interactions that seem well suited to confer specificity. All side chains that make direct contact with the target sequence are presented by the aptly named recognition helix. A characteristic of these interactions is that they are interwoven: a protein side chain makes contact with more than one base pair, and the identity-determining base pairs in turn interact with more than one side chain. These interlocking interactions are probably the basis for the high degree of conservation observed in both the target DNA sequence and the E2 protein.

Although the structures of protein–DNA complexes are beautiful to look at, attempts to codify the underlying interactions are decidedly problematic. The first high-resolution crystal structures of such complexes were reported four years ago, and the dozen or so structures we now have in hand suggest that, like protein folding, the problem of DNA recognition will not be easily understood[3]. The elegant simplicity of information storage in DNA is not reflected in the structures of proteins that recognize DNA: nature's rule for designing these proteins seems to be that anything goes. The particular interactions between protein side chains and DNA base pairs depend in a highly complex way on the conformation of the entire assembly, and are not simply dictated by the DNA sequence alone.

Roles

Likewise, the particular roles played by protein and DNA structural elements differ from case to case. For example, E2 uses an α-helix to present side chains for interactions with DNA base pairs. In this case, the α-helix runs roughly parallel to the major groove. Other protein–DNA complexes show a range of orientations for the α-helix, and also show different parts of the helices providing contacts with DNA[3]. Another class of transcription factors uses pairs of β-strands that bind to the major groove of DNA[5], while there are yet others, such as the TATA-box-binding protein (for review see ref. 6), that appear to make specific contacts with the minor groove.

The determination of the three-dimensional structure and the resulting tabulation of hydrogen bonds and other interactions is just the starting point for understanding the physical chemistry of DNA-binding specificity, which must take into account interactions with water molecules, counterions and with non-specific DNA sequences. Unravelling these details will keep theoreticians busy for a while yet.

In the meantime, we look forward to augmenting this rich harvest of DNA-binding modules with structural information about the equally important activation domains responsible for communication with other elements of the transcription apparatus. Such domains have not yet yielded to structural analysis, perhaps because they are less likely to be stable enough to make the application of crystallography and NMR straightforward. Ingenuity, as well as the application of other physical techniques, will be required to work out a complete picture of transcription factor action. □

John Kuriyan and Stephen K. Burley are in the Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

1. Hegde, R. S., Grossman, S. R., Laimins, L. A. & Sigler, P. B. *Nature* **359**, 505–512 (1992).
2. Harrison, S. C. *Nature* **353**, 715–719 (1991).
3. Pabo, C. O. & Sauer, R. T. *A. Rev. Biochem.* **61**, 1053–1095 (1992).
4. Ham, J. *et al. Trends biochem. Sci.* **16**, 440 (1991).
5. Phillips, S. E. V. *Curr. Opin. struct. Biol.* **1**, 89–98 (1991).
6. Sharp, P. A. *Cell* **68**, 819–821 (1992).

© 1992 Nature Publishing Group