nature structural biology
March

# Feeling the groove

**Variation in ways in which proteins can identify and lock on to genes due for transcription is illuminated by studies of the Arc repressor, which has a $\beta$-sheet, rather than $\alpha$-helical, recognition motif.**

CENTRAL to any data storage system is the ability to access and retrieve information. Each block of data must be identified to distinguish it from the other material in the data base. The nature of the retrieval system is then determined by the data storage medium. The information stored in a genome, in the form of gene sequences, must be accessed, often in a highly selective manner, and converted into the machinery that builds, and maintains, the organism in question.

Proteins of the transcription apparatus, the nuts and bolts of the genome information retrieval system, are able to identify the particular gene or genes that are destined to be transcribed into RNA, which is then frequently translated into protein. New studies[1,2] of one such transcription factor, the Arc repressor of *Salmonella* bacteriophage P22, provide an illustration of the means by which a protein can identify and get to grips with the short DNA sequence that flags the presence of a particular gene.

Consideration of the structure of DNA indicates that there is enough information provided by the hydrogen bond donors and acceptors on the exposed edges of the bases in the major, but not the minor, groove to distinguish each of the four bases in DNA[3]. Indeed, most sequence-specific recognition of DNA is through the major groove.

What kind of protein motifs 'feel the groove'? Early studies noted that a two-stranded anti-parallel $\beta$-sheet has features that are complementary to those of both double-stranded RNA[3] and DNA[4] and such peptides fit comfortably into the minor and major grooves of DNA[4–6]. In contrast the first structures of sequence-specific DNA-binding proteins, all prokaryotic regulators, effected sequence recognition in the major groove by a so-called recognition $\alpha$-helix[7].

To date, most of the known sequence-specific interactions with DNA are mediated by residues on $\alpha$-helices. These $\alpha$-helices are presented to the major groove in the context of a variety of structural elements (the helix-turn-helix, the basic–helix-loop-helix and so on)[7]. Nonetheless $\beta$-sheet motifs are also used in proteins that interact with both the major and, surprisingly, minor groves.

Also in this month's *Nature Structural Biology:* folding of cytochrome *c* in the absence of intermediates; dimerization of transmembrane $\alpha$-helices; two distamycin A molecules in the minor groove of B-DNA; the mechanism of allosteric activation in bacterial L-lactate dehydrogenase; substrate-channelling in dihydrofolate reductase-thymidylate synthase; and growing protein crystals on lipid layers.

The Arc repressor is a member of the ribbon-helix-helix family of transcription factors, which also includes the MetJ repressor[8]. Both proteins bind to their cognate operator as tetramers, consisting of a dimer of dimers. Each monomer provides one of the ribbons of the recognition $\beta$-sheet and each dimer interacts with an operator 'half-site', TAGA in the case of Arc. Although individual Arc dimers display two-fold symmetry, the operator half-site to which they bind is asymmetrical. Inevitably, symmetry-related amino acids must interact with non-symmetry-related base-pairs; Asn 11' on one $\beta$-ribbon interacts with T(A:T)GA in the TAGA box whereas Asn 11 on the other $\beta$-strand makes a contact with TA(G:C)A. It seems likely that the flexibility of the amino-acid side chains at the interface facilitate this asymmetrical interaction.

The different DNA contacts mediated by equivalent residues in the Arc dimer illustrates the scope for variability in protein–DNA interactions and again emphasizes the difficulties of trying to decipher any type of detailed code for protein–DNA interactions.

An important part of the recognition process seems to involve a conformational adjustment of the $\beta$-sheet on binding; two phenylalanines are unpacked from the hydrophobic core and are inserted between phosphate oxygens of the DNA backbone, for example. Overall, the interaction of the Arc dimer with the phosphate backbone lining the edges of the major groove and flanking the $\beta$-ribbon is symmetrical. Such interactions are likely to lock the $\beta$-sheet tightly into the major groove, preventing wobbling or twisting of the sheet.

The mutagenesis experiments emphasize that although the recognition sheet furnishes the base-specific contacts, effective sequence recognition depends on a range of different and not necessarily direct interactions between DNA and protein. Base-specific interactions are essential, of course. Surprisingly, asymmetrical rather than symmetrical contacts contribute most to binding energy. But interactions across the tetramer interface, involved in cooperativity, and linkage of $\beta$-sheet–major groove to phosphate contacts are also important for binding.

Two other classes of DNA-binding protein have been characterized that interact with DNA through $\beta$-sheet motifs. The prokaryotic HU family of proteins[8] is thought to use a two-stranded $\beta$-sheet to interact with the minor groove, in a fashion that is reminiscent of the earlier modelling studies[4,5]. As might be expected, given the paucity of sequence-dependent features in the minor groove, the HU protein binds DNA non-specifically. Surprisingly the IHF protein binds in sequence-specific manner. How this is achieved will have to await the structure of the protein–DNA complex.

The 'saddle and stirrups' binding surface of the TATA-box binding protein TBP, which consists of a 10-stranded sheet, also interacts in a sequence-specific manner with the minor groove. The structure of the complex reveals a highly unusual protein–DNA interface[9,10]. The DNA is grossly distorted and the minor groove is splayed open, permitting a series of side-chain–base interactions that, in combination with the deformability of the run of TATA bases, determine sequence specificity.

Why is the $\alpha$-helix the most prevalent recognition motif used by sequence-specific DNA-binding proteins? Is there some intrinsic advantage in using this motif, as opposed to $\beta$-sheets? Perhaps the $\beta$-sheet fulfils a different role — as in TBP? Or are the features of present-day DNA-binding proteins merely a reflection of the vagaries of evolution? At least some of these questions should be answered by the ever more detailed study of this fascinating class of proteins.

**Guy Riddihough**

*Guy Riddihough is Editor of* Nature Structural Biology.

1. Raumann, B.E. *et al. Nature* **367**, 754–757 (1994).
2. Brown, B.M.*et al. Nature struct. Biol.* **1**, 164–168 (1994).
3. Seeman, N.C., Rosenberg, J.M. & Rich, A. *Proc. natn. Acad. Sci. U.S.A.* **73**, 804–808 (1976).
4. Carter Jr, C.W. and Kraut, J. *Proc. natn. Acad. Sci. U.S.A.* **71** 283–287 (1974).
5. Church, G.M., Sussman, J.L. & Kim, S.-H. *Proc. natn. Acad. Sci. U.S.A.* **74**, 1458–1462 (1977).
6. Chou, P.Y., Alder, A.J. & Fasman, G.D. *J. molec. Biol.* **96**, 29–45 (1975).
7. Pabo, C.O. & Sauer, R.T. *A. Rev. Biochem.* **61**, 1053–1095 (1992).
8. Phillips, S.E.V. *Curr. Opin. struct. Biol.* **1**, 89–98 (1991).
9. Kim, Y. *et al. Nature* **365**, 512–520 (1994).
10. Kim, J.L., Nikolov, D.B. & Burley, S.K. *Nature* **365**, 520–527 (1994).