*compounds* which recognize short DNA sequences specifically. These attempts often start from the X-ray crystal structures of complexes between short pieces of DNA and various antibiotics, or chemical compounds such as ethidium or porphyrin. In another important kind of study, scientists have tried to design novel peptides, or peptide-like compounds, that will bind specifically to base-pairs within either the major or the minor groove of DNA. There are many natural examples of this sort, which show useful biological activities, including the antibiotics actinomycin, daunorubicin, chromomycin and bleomycin – some of which are still used for the treatment of cancer.

As the best known example, Peter Dervan and colleagues have synthesized a series of distamycin-related compounds, which place an amide chain deeply into the minor groove of DNA, so as to recognize base-pairs there. The native distamycin recognizes a short base sequence such as AAAA or AATT, by means of hydrogen bonds from N–H groups on the amides to nitrogens N or oxygens O on the minor-groove edges of A–T base-pairs . But Dervan and his chemists have gone one step further, and have made polyamides which place *two* chains in the minor groove, lying side-by-side, as shown schematically in Fig. 8.9(a). Those synthetic poly-amides may contain separate ring units of pyrrole (Py), imidazole (Im) and hydroxy-imidazole (Hp). Pyrrole is the five-membered ring compound found in the red pigment of blood, while imidazole is part of the side chain of the amino acid histidine. Two such polyamide chains can be linked to form a kind of hairpin, so that the Py, Im and Hp stack as pairs. Those amide units may bind to the minor-groove oxygen and nitrogen atoms by means of hydrogen bonds; and it turns out that Py and Im, side-by-side in the minor groove, will recognize uniquely a C–G base-pair, as shown schematically in Fig. 8.9(b) and (c). Similarly, Py and Hp side-by-side will recognize uniquely an A–T base-pair, as also shown. Those poly-amides can rotate around the bonds that link their amide units together; and so they can coil helically in space to maintain close register with base-pairs in the helical minor groove.

To conclude our survey, we have seen here how different proteins can recognize specific base sequences of DNA by many different mechanisms, most of which are not predicted by current theory. Most of those proteins seem first to 'dock' in a rough fashion onto the sugar–phosphate chains of DNA, so as to insert an α-helix or β-sheet into the major groove; then in a second step, the amino acids which protrude into the groove may 'probe' for the identities of base-pairs, by means of hydrogen bonds or hydrophobic contacts. Such precise interactions cannot easily be predicted by theory,

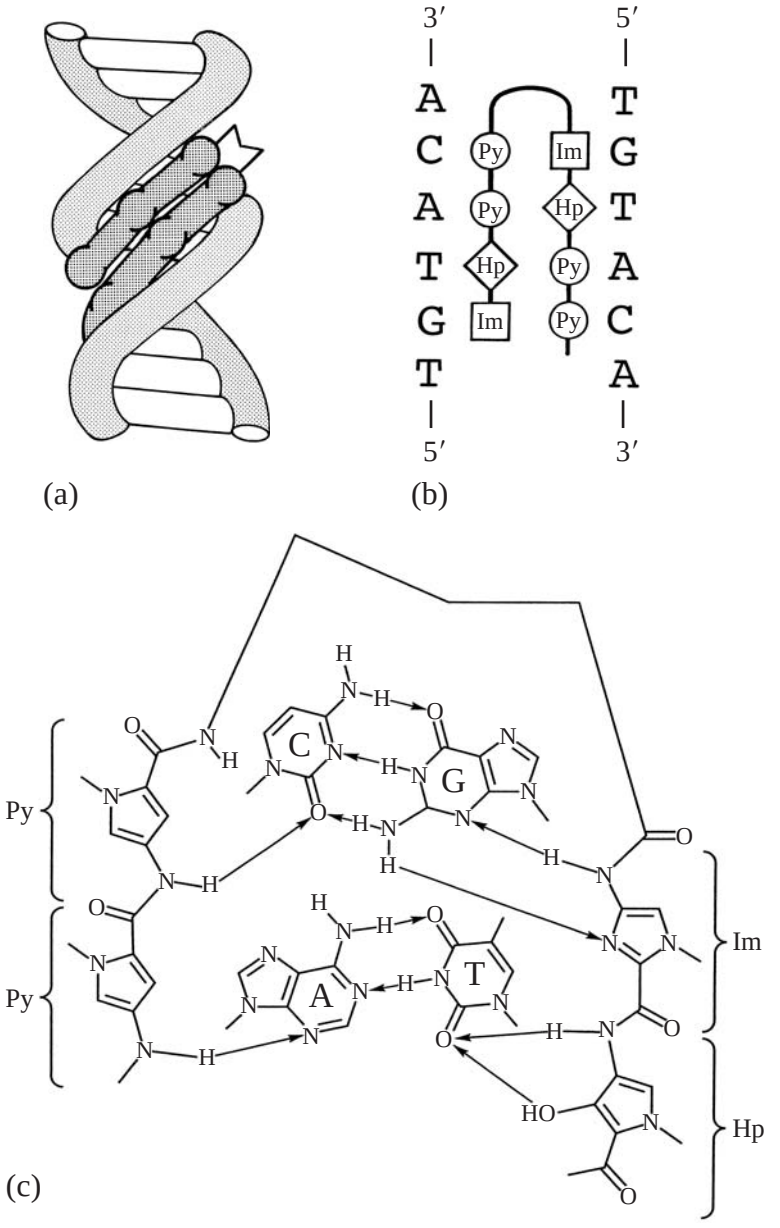



**Figure 8.9** Recognition of specific base sequences by side-by-side poly-amide molecules in the minor groove. (a) General arrangement, showing two chains linked at one end. (b) Spread-out schematic view of the minor groove of a particular sequence, with the two amide chains in register with their recognition partners. This shows the unique recognition code. (c) Detailed layout of the hydrogen bonding for the recognition of C–G and A–T base pairs. Here, the poly-amides are shown on either side of the base-pairs: the poly-amides have been separated and spread out on the page, while the base-pairs have been rotated onto the same plane. The scales of the poly-amides and the bases are not the same. Hydrogen bonds are shown by arrows, in the direction of donation.

because they involve intricate stereochemistry. Moreover, the DNA helix is flexible in a way that depends on its base-sequence; which makes prediction even more difficult.

Nevertheless, some workers have made progress in this difficult field, by selecting novel DNA-binding sites for zinc-finger proteins, through random mutation of amino acids in a single zinc-finger, which then binds with altered specificity to the DNA. Other workers have made novel chemicals which bind to DNA of specific base sequences; and those small molecules may also be useful in biology and medicine. Studies of the specific interaction between proteins and DNA are still at an intermediate stage; and no doubt more progress will be made in the next ten to twenty years, which should result in some useful inventions.

The cellular biology of specific protein–DNA interactions remains far beyond the scope of our limited text; but a few points about such biology may be made briefly. First, one might imagine that every gene in a complex organism would have its own particular regulatory protein, say a repressor or activator, which could bind specifically to some particular DNA sequence near the start-site for transcription of the gene. Given 30 000 genes in the human genome, such a one-to-one scheme would require 30 000 unique regulatory proteins! Yet Nature often does things differently from ways that one might naively imagine.

Thus, in most kinds of bacteria, several related genes may be clustered together, using only *one* protein to control their overall expression. A gene-cluster of that kind, known as an 'operon', may contain most or all of the necessary genes for some particular biochemical process: say all of the enzymes needed for synthesis of the amino-acid methionine. The bacterial methionine operon is in fact controlled by just one protein: the *met* repressor that we described earlier in this chapter.

Now in animals or plants, several related genes may still cluster together; yet due to the great complexity of a large genome, one protein is usually not enough to control transcription with sufficient specificity. And so we find that most genes in animals or plants are controlled by two or more proteins known as 'transcription factors', which can act jointly in various combinations so as to achieve a much higher gene-specificity than for any single protein acting separately.

Furthermore, the cells of animals or plants contain very complex DNA-packaging devices in the form of chromosomes, which also regulate genes in specific ways. Thus, the histone proteins which condense very long DNA into chromosomes may be modified chemically by enzymes known as 'acetylases' or 'methylases', so as

to reduce the tightness of packing around specific genes, and thereby activate transcription in particular locations (see Chapters 7 and 11). The genes of animals or plants, therefore, would seem to be controlled by a complex hierarchy of molecular interactions; whereas the genes of bacteria may often be controlled more simply.

## Further Reading

Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M., and Harrison, S.C. (1988) Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**, 899–907. An early example of specific protein-DNA binding, for the 434 repressor bound to its preferred base sequence. Source of Figs 8.2(a) and 8.3.

Arora, P.S., Ansari, A.Z., Best, T.P., Ptashne, M., and Dervan, P.B. (2002) Design of artificial transcriptional activators with rigid poly-L-proline linkers. *Journal of American Chemical Society* **124**, 13067–71. How polyamide based compounds may be used to control gene expression.

Fersht, A.R. (1999) *Structure and Mechanism in Protein Science*. W.H. Freeman, New York. Elucidates the general principles of specific associations within and between biological molecules, including interactions that establish the fold of a protein, and the protein-to-protein interactions that affect their stability and specificity.

Gowers, D.M. and Halford, S.E. (2003) Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling. *EMBO Journal* **22**, 1410–18. How proteins can find target sites on a length of DNA by sliding or hopping. These processes influence the rates at which sites can be found within the cell.

Judson, H.F. (1979) *The Eighth Day of Creation*. Simon & Schuster, New York. Chapter 2(a) provides an excellent description of Pauling's early studies of proteins, and of his discovery of the simple α-helix structure.

Lewis, M., Chang, G., Horton, N., Kercher, M., Pace, H., Schumacher, M., Brennan, R., and Lu, P. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271**, 1247–55. A well-known repressor protein bound to DNA; two such repressor proteins often stick together, to form a tight DNA loop.

Li, T., Stark, M., Johnson, A., and Wolberger, C. (1995) Crystal structure of the MATa1/MATa2 homeodomain heterodimer bound to DNA. *Science* **270**, 262–7. A DNA-binding protein from yeast, which curves the DNA around itself into a large loop. The structure also illustrates how protein-to-protein interactions are an important aspect of specificity and control of gene expression.

Lilley, D.M.J. and White, M.F. (2001) The junction-resolving enzymes. *Molecular Cell Biology* **2**, 433–43. A summary of proteins which recognize 'four-way junctions' in DNA to facilitate homologous recombination.

Marmorstein, R. and Fitzgerald, M.X. (2003) Modulation of DNA-binding domains for sequence-specific DNA recognition. *Gene* **304**, 1–12. Factors

that influence DNA sequence recognition, including protein-to-protein interactions in multi-component assemblies of DNA-binding proteins.

Nolte, R.T., Conlin, R.M., Harrison, S.C., and Brown, R.S. (1998). Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proceedings of the National Academy of Sciences, USA* **95**, 2938–43. Example of different modes of DNA binding by the zinc-finger family.

Ogata, K., Sato, K., and Tahirov, T. (2003) Eukaryotic transcriptional regulatory complexes: cooperativity from near and afar. *Current Opinion in Structural Biology* **13**, 40–8. Principles of combinatorial control of gene regulation in eukaryotes; experimental evidence for DNA looping in a multi-component regulatory complex, using atomic force microscopy – a technique described in Chapter 9.

Perutz, M.F. (1992) *Protein Structure: New Approaches to Disease and Therapy*. W.H. Freeman, New York. An authoritative survey of protein structures, and their relevance to drug design and medicine.

Rice, P.A., Yang, S., Mizuuchi, K., and Nash, H.A. (1996) Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87**, 1295–306. The remarkable structure of the IHF protein bound to DNA, which includes 180 degrees of curvature over 35 base pairs, and proline intercalation at two large kinks.

## Web-based Resources

Atlas of amino-acid/base interactions by Janet Thorton and colleagues:
**http://www.biochem.ucl.ac.uk/bsm/sidechains**

A summary of DNA-binding protein structural families
**http://www.biochem.ucl.ac.uk/bsm/prot_dna/prot_dna_cover.html**

A repository of crystallographic and nuclear magnetic resonance spectroscopy structures of nucleic acids and protein complexes
**http://www.ndbserver.rutgers.edu**

## Bibliography

Choo, Y. and Klug, A. (1997) Physical basis of a protein-DNA recognition code. *Current Opinion in Structural Biology* **7**, 117–25. Describes studies on the engineering of specificity of DNA binding protein, with emphasis on the zinc-fingers for base triplets in DNA, when amino acids are altered in key positions. A source for Fig. 8.6.

Dervan, P.B. (2001) Molecular recognition of DNA by small molecules. *Bioorganic Medicinal Chemistry* **9**, 2215–35. An account of the development of the poly-amide compounds, based on observations of small molecules binding to DNA. The source of Fig. 8.9.

Jamieson, A.C., Kim, S.-H., and Wells, J.A. (1994) *In vitro* selection of zinc fingers with altered DNA-binding specificity. *Biochemistry* **33**, 5689–95.

Changes of amino acids within the zinc-fingers of Zif268 allow them to recognize different DNA sequences.

Keller, W., König, P., and Richmond, T.J. (1995) Crystal structure of a bZIP/DNA complex at 2.2 Å resolution: determinants of DNA specific recognition. *Journal of Molecular Biology* **254**, 657–67. Crystal structure of the bZIP protein bound to DNA. Source of Fig. 8.4.

Kielkopf, C.L., White, S., Szewczyk, J.W., Turner, J.M., Baird, E.E., Dervan, P.B., and Rees, D.C. (1998) A structural basis for recognition of A-T and T-A base pairs in the minor groove of B-DNA. *Science* **282**, 111. Crystal structure showing how a poly-amide can encode recognition of all four possible base-pairs (A–T, T–A, G–C and C–G) through minor groove contacts. Source of Fig 8.8.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001) Amino acid-base interactions: a three-dimensionsal analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research* **29**, 2860–74. A review of amino acid-to-base contacts in a wide variety of protein-DNA complexes.

Miller, J.C and Pabo, C.O. (2001) Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *Journal of Molecular Biology* **313**, 309–15. Explores detailed issues concerning the design of zinc-finger proteins to recognize specific DNA targets. A source for Fig. 8.6.

Nair, S.K. and Burley, S.K. (2003) X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular basis of regulation by proto-oncogenic transcription factors. *Cell* **112**, 193–205. Illustrates how DNA is recognized by one class of zipper proteins, and shows how the Myc-Max heterodimer could favor loop formation in the DNA.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **335**, 321–9. Most of the contacts between protein and DNA bases are mediated indirectly by water in this structure.

Pabo, C.O. and Nekludova, L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *Journal of Molecular Biology* **301**, 597–624. Demonstrates geometrical requirements for favorable hydrogen-bonding interactions between the DNA bases and amino-acid side chains, and how they depend on the orientation of α-helices in the groove.

Pavletich, N.P. and Pabo, C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809–17. A chain of three zinc-fingers that bind to DNA in a regular way, for the protein Zif268. A source of Fig. 8.6.

Rastinejad, F. and Khorasanizadeh, S. (2001) Nuclear-receptor interactions on DNA-response elements. *Trends in Biochemical Scences* **26**, 384–90. Summarizes crystal structures of different receptor proteins on different DNA half-sites, and recognition of symmetry through protein-to-protein interactions. A source of Fig. 8.7.

Schaffer, P.L. and Gewirth, D.T. (2002) Structural basis of VDR-DNA interactions on direct repeat response elements. *EMBO Journal* **21**, 2242–52. A source of Fig. 8.7.

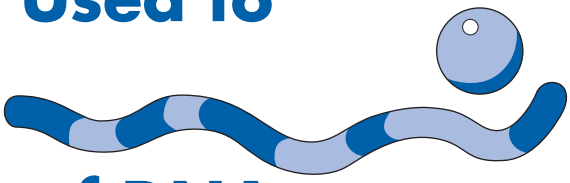
Somers, W.S. and Phillips, S.E.V. (1992) Crystal structure of the *met* repressor-operator complex at 2.8 Å resolution reveals DNA recognition by β-strands. *Nature* **359**, 387–93. The first detailed example of a β-sheet structure which binds specifically to DNA. Source of Fig. 8.5(a) and (c).

Suzuki, M. and Yagi, N. (1996) An in-the-groove view of DNA structures in complexes with proteins. *Journal of Molecular Biology* **255**, 677–87. Studies of the detailed fit between transcription factor proteins and DNA.

Tsai, F.T., Littlefield, O., Kosa, P.F., Cox, J.M., Schepartz, A., and Sigler, P.B. (1998) Polarity of transcription on PolII and archael promoters: where is the 'one-way sign' and how is it read? *Cold Spring Harbor Symposium in Quantitative Biology* **63**, 53–61. A source of Fig. 8.7.

# CHAPTER 9

# Methods Used to Study the Structure of DNA



Our goal in this book has been to explain, as simply as possible, how DNA works in biology. For that reason, we have tried not to dwell too much on the methods which are used by scientists to study DNA: instead we have tried to give an integrated picture of DNA as obtained by many different methods of analysis. We have emphasized on many occasions that DNA is a very tiny object; yet our pictures of DNA have been drawn in terms of images which may be perceived by the reader on a 'household' scale.

A student who wants to understand any subject in depth will want to know exactly how the evidence has been obtained, from which the overall conclusions have been reached. And the historian of whom we spoke in Chapter 1 was puzzled not so much by the fact that the DNA in the cells of our bodies is so exceedingly small, but by the problem of how one can *find out* anything about an object so small. In this chapter, therefore, we shall explain some of the techniques which scientists today use to study the structure of DNA.

The most important method, at least from a historical point of view, has been the analysis of DNA structure by *X-ray diffraction*. This is the tool which was used to discover the basic double-helical form of DNA in 1953. Ten years earlier, in the 1940s, studies of pneumococcal bacteria by Oswald Avery and colleagues had shown that a pure preparation of DNA could cause a harmless form of the bacterium to become infectious, and so impart pneumonia to mice. (We know now that the DNA used by Avery contained a gene for making a strong shell or coat around the bacterium, but this was not known at the time.) By the 1950s, enough evidence had piled up to convince even physical scientists that DNA might constitute the invisible 'gene' of which biologists had spoken for more than 40 years. Therefore, some

physicists and chemists began to investigate the structure of DNA by various methods, including X-ray diffraction, to see whether its physical structure might shed any light on how DNA could act as the genetic material.

A talented early X-ray worker was Rosalind Franklin. She pulled fine fibers of DNA from natural sources, and found that when those fibers were exposed to X-rays, they could give either of two distinct X-ray diffraction photographs. She called these two patterns 'A' and 'B'. The 'A' form was seen when she kept the fibers relatively dry, whereas the 'B' form was seen when she kept the fibers wet. Her 'B' form photograph – which was much the simpler of the two – was interpreted by James Watson and Francis Crick in the spring of 1953 in terms of a right-handed double helix containing A–T and G–C base-pairs. Robert Langridge subsequently confirmed the essential points of the 'B' form model in 1960; and Watson Fuller produced a similar model for the 'A' form in 1965, refining a model first proposed by Franklin and Raymond Gosling late in 1953. These 'A' and 'B' form models were shown as part of Fig. 2.7.

Later work by Struther Arnott and colleagues showed that DNA of a regularly repeating sequence, such as $A_n/T_n$ (that is, all A on one strand and all T on the other) or $(AT)_n/(AT)_n$ (that is, the alternating sequence ATATAT on both strands) could be extremely polymorphic. Thus, each fiber preparation could produce as many as three or four different kinds of X-ray pattern, depending on its salt and water content during exposure to X-rays. For example, a fiber of the sequence $(AT)_n/(AT)_n$ can produce a total of four different X-ray patterns under different conditions, which are known as 'A', 'B', 'C', and 'D'. Other base sequences can produce related X-ray patterns, for example 'B'', 'C'', and 'C''', which are clearly variants of the basic forms; while still others show X-ray patterns such as 'E' that are plainly distinct. Thus by 1980, much evidence had accumulated that the structure of DNA might be more complex than Watson and Crick could ever have anticipated in the 1950s. Yet although X-ray pictures of a fiber sample can show well enough that the forms 'A' to 'E' are distinct, they do not yield enough information to determine the three-dimensional structures of those different forms at sufficient resolution to see the individual atoms clearly.

Fortunately, by 1980, chemists such as Keiichi Itakura, Shoji Tanaka, and Jacques van Boom had learned how to synthesize DNA chemically in large amounts, and how to purify it so that one could grow crystals of particular, short base sequences. Crystals will not grow unless the preparation is pure; that is, unless the short fragments of DNA (or oligomers,[1] as they are known) have identical sequence and are all of the same length. The first structure of DNA to

be solved by X-ray analysis of a crystal, as distinct from a fiber, was that of the sequence ATAT by M.A. Viswamitra in 1978. It proved to be disappointing, for the molecule did not form a complete double helix, perhaps because the TA step unwinds easily. But the next few X-ray structures were to produce astonishing results: the sequences CGCG and CGCGCG, as analysed independently by Andrew Wang, J. Crawford, Horace Drew, and their colleagues in 1979–80, both crystallized as left-handed double helices. Biologists had assumed for over 20 years that DNA could only be right-handed; and then it was discovered that DNA could be left-handed as well! Earlier solution studies by Fritz Pohl and Tom Jovin, using circular dichroism methods (see below), had suggested that alternating C–G sequences such as CGCG might be either right-handed or left-handed, depending on the salt concentration; but only a few crystallographers and other specialists had taken them seriously.

Since 1980, many important structures of DNA oligomers, and of their complexes with antibiotics or proteins, have been analysed in single crystals by X-ray diffraction methods. These studies form the essential background for all of the previous chapters in this book. They are now far too numerous to be cited here, but we list some references to this large body of work at the end of the chapter. Here we shall explain the typical steps of an X-ray analysis with reference to a particular DNA molecule, or protein–DNA complex of interest.

First, the crystallographer must decide what sequence of DNA to study, and then prepare large amounts of the material in pure form, usually by the method of chemical synthesis. In the case of a protein–DNA complex, one has to prepare also large amounts of the protein in chemically pure form, usually by cloning a gene for the protein into bacteria, and then expressing the protein in large amounts. Next, he or she must grow crystals of this substance that are suitable for X-ray diffraction studies. Growing crystals is a very chancy business! Some crystals of DNA, suitable for X-ray analysis, are shown in Fig. 9.1. They are each about 1 mm long. These particular crystals contain a complex of an eight-base-pair double helix of sequence AGCATGCT, together with the antibiotic nogalamycin to which it binds tightly. The antibiotic is orange; the DNA itself has no color. The crystals shown in this figure were grown slowly in a cold room for 2 weeks before they reached the required size of 1 mm.

Next, one of the crystals is carefully mounted in a wet, sealed capillary tube and placed in an X-ray beam. If the crystal is well-ordered in structure, an X-ray photograph such as that shown in Fig. 9.2 will be obtained. Any large crystal of DNA is made from millions of identical DNA molecules, all of which are close-packed into some sort of regular array. The geometrical locations of spots in Fig. 9.2 tell us