Review

An overview of the structures of protein-DNA complexes

Nicholas M Luscombe*, Susan E Austin*, Helen M Berman† and Janet M Thornton*‡

Addresses: *Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK. †Department of Chemistry, Rutgers State University, Piscataway, New Jersey 08855, USA. ‡Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK. E-mail: nick@biochem.ucl.ac.uk; s.austin@biochem.ucl.ac.uk; berman@rcsb.rutgers.edu; thornton@biochem.ucl.ac.uk

Correspondence: Janet M Thornton.

Published: 9 June 2000

Genome ***Biology*** 2000, **1(1)**:reviews001.1–001.37

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/1/reviews/001

© Genome**Biology**.com (Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

On the basis of a structural analysis of 240 protein-DNA complexes contained in the Protein Data Bank (PDB), we have classified the DNA-binding proteins involved into eight different structural/functional groups, which are further classified into 54 structural families. Here we present this classification and review the functions, structures and binding interactions of these protein-DNA complexes.

Introduction

DNA-binding proteins have a central role in all aspects of genetic activity within an organism, such as transcription, packaging, rearrangement, replication and repair. It is therefore extremely important to examine the nature of complexes that are formed between proteins and DNA, as they form the basis of our understanding of how these processes take place. Over the past ten years, we have witnessed a great expansion in the determination of high-quality structures of DNA-binding proteins. The structures, especially those of their complexes with DNA, have provided valuable insight into the stereochemical principles of binding, including how particular base sequences are recognized and how the DNA structure is quite often modified on binding.

A classification of protein-DNA complexes based on the structures of the DNA-binding regions in the proteins is described. The taxonomy was first proposed by Harrison [1] and later modified by Luisi [2]. Here, we build on the original classification with appropriate extensions to accommodate the new structures that have been solved. Assembling the structures in such a system simplifies comparison of the different modes of binding, allowing identification of common themes between structurally related proteins and also highlighting unusual features that distinguish a particular protein from others. Examination of genes that are functionally assigned in PEDANT [3] show that typically 2-3% of a prokaryotic genome and 6-7% of a eukaryotic genome encodes DNA-binding proteins. Therefore, the classification of structures presented here is far from complete and many more entries are anticipated. It should be noted that the number of structures in the PDB does not necessarily reflect the relative importance of the protein in the organism. The helix-turn-helix (HTH), the ββα zinc finger, and the zipper-type motifs are, however, expected to be very common.

This review provides a general overview of the DNA-binding structures that have been found, along with detailed descriptions of the individual protein families. As our main research interest lies in the investigation of interactions between proteins and DNA, the main focus is on X-ray structures of complexes that provide the requisite details. Also introduced is a new web-based resource of protein-DNA complex structures.

Constructing the classification

Dataset of protein-DNA complex structures

Protein-DNA complexes solved by X-ray crystallography to a resolution of higher than 3.0 Å were obtained from the January 2000 release (04/01/00) of the Protein Data Bank (PDB) [4,5] and the Nucleic Acid Database (NDB) [6]. The complexes were defined as any structure containing one or more protein chains and at least one double-stranded DNA of more than four base-pairs (bp) in length. From this set we excluded structures containing single- and quadruple-stranded DNA and non-contiguous DNA (that is, with a break in the strand). This resulted in a dataset of 240 protein-DNA complexes (Tables 1,2). Box 1 shows the selection process.

Included in the dataset were 24 homodimeric complexes whose asymmetric unit contained only half the structure. The full coordinate files were obtained from the NDB, which calculates the coordinates for the complete molecule by applying the transformation matrices provided in the PDB files to the half structure. These entries are marked accordingly in Table 2.

Structural taxonomy and classification of protein-DNA complexes

The PDB entries were classified according to the structures of the proteins in the complex. The classification system categorizes them in a two-tier system at the group and family levels. At the first level, proteins were sorted manually into eight groups by visual inspection using RasMol [7] and from the literature. Members of the same group share a prominent structural feature used for DNA recognition and are related to each other in varying degrees. The eight groups are: (I) HTH (including 'winged' HTH), (II) zinc-coordinating, (III) zipper-type, (IV) other α helix, (V) β sheet, (VI) β hairpin/ribbon, (VII) other, and (VIII) enzymes (Table2). The enzyme group is an exception to the structural criterion, as it contains all proteins that display enzymatic activity when bound to DNA. Five enzymes also qualify on structural grounds for the HTH and 'other α helix' groups: restriction endonuclease *Fok*I (PDB entry 1fok), γδ-resolvase (1gdt), Hin recombinase (1hcr), Tc3 transposase (1tc3) and Cre recombinase (1crx). These proteins are listed under the HTH group in Table 2 and are marked appropriately.

At the second level of classification, the DNA-recognition domains were classified into homologous families by comparing their structures in pairs using the secondary structure alignment program SSAP [8]. The program uses a dynamic programming method [9] and assesses the similarity between proteins by comparing the structural environments of the constituent amino acids. SSAP returns a score of 100 for identical proteins, and >80 for homologous proteins; proteins are automatically assigned to the same family if they score above this cut-off. More distantly related proteins that give scores of >70 are also placed in the same family if they perform similar biological functions [10].

Proteins were broken down into their constituent DNA-binding domains before conducting the alignments. In most dimers, each domain corresponds to a distinct subunit and the structure simply needs to be separated into the constituent chains. In proteins such as those with ββα zinc fingers, however, a chain contains several binding domains; in such cases, therefore, the subunits were separated into the appropriate segments, which are listed in Table 2. In this review, structures are identified by the standard four-digit PDB code (for example, 1aay).

Table 1

The groups of protein structures found in the dataset, the number of families within each group and the number of PDB files each family contains.

		Nº. of proteins (PDB files) in group			
Protein group	Nº. families in group	Prokaryote	Eukaryote	Viral	Total
1. Helix-turn-helix	16*	32	28	-	60
2. Zinc-coordinating	4	0	23	-	23
3. Zipper-type	2	0	10	-	10
4. Other α-helix	7	1	5	2	8
5. β-sheet	1	0	8	-	8
6. β-hairpin/ribbon	6	10	1	-	11
7. Other	2	0	8	-	8
8. Enzyme	16	43	68	2	113
Total	54	86	151	4	241†

*Includes the two 'winged' helix-turn-helix families.
†PDB entry 1a02 contains proteins belonging to the families 'zipper-type' and 'other'.

Table 2

List of the 240 structures of protein-DNA complexes in the dataset.

PDB code	DNA-binding subunits	Protein name	Species	Resolution (Å)	DNA sequence
I. Helix-turn-helix group					
1. Cro and Repressor family					
1lmb*	3,4	Repressor	Phage λ	1.8	-AATACCACTGGCGGTGATATTATAT-CACCGCCAGTGGTAT-
1lli	A,B	Repressor mutant	Phage λ	2.1	-AATACCACTGGCGGTGATATTATAT-CACCGCCAGTGGTAT-
1per	L,R	Repressor	Phage 434	2.5	AAGTACAGTTTTTCTTG-TATTATA--CAAGAAAAACTGTACT
1rpe	L,R	Repressor	Phage 434	2.5	-TATACAATGTATCTTG-TTTGACAAACAAGATACATTGTAT-
2or1	L,R	Repressor	Phage 434	2.5	AAGTACAAACTTTCTTG-TATTATA--CAAGAAAGTTTGTACT
3cro	L,R	Cro	Phage 434	2.5	AAGTACAAACTTTCTTG-TATTATA--CAAGAAAGTTTGTACT
6cro	A	Cro	Phage λ	3.0	AAGTACAAACTTTCTTG-TATTATA-CAAGAAAGTTTGTACT
3orc	A	Cro	Phage λ	3.0	AAGTACAAACTTTCTTG-TATTATA--CAAGAAAGTTTGTACT
2. Homeodomain family					
1fjl*†	A,B,C,G	Paired protein	*D. melanogaster*	2.0	-----AATAATCTGATTACTGTAATCAGATTAT----------
1hdd	C,D	Engrailed homeodomain	*D. melanogaster*	2.8	--TTTTGCCATGTAATTACCTAAATTAGGTAATTACATGGCAAA
1apl	C,D	Mat α–2	*S. cerevisiae*	2.7	--ACATGTAATTCATTTACACGCTGCGTGTAAATGAATTACATG
1yrn	A,B	Mat a-1/α–2	*S. cerevisiae*	2.5	-TACATGTAATTTATT-ACATCATATGATGTAATAAATTACATG
1au7	A1(5-76), A2(103-160), B1(5-75), B2(102-161)	Pit-1 POU domain	*R. norvegicus*	2.3	-----TTGCGTAGCGTTACGTATATTCCGCTAATCGAT------
1oct	C1(5-75) C1(102-161)	Oct-1 POU domain	*H. sapiens*	3.0	----------TGTATTGCAAATAAGGACCTTATTTGCATAC---
2hdd	A,B	Engrailed homeodomain	*D. melanogaster*	1.9	----------TGTTTTTGATAAGATCTTATCAAAAAC-------
3hdd	A,B	Engrailed homeodomain	*D. melanogaster*	2.2	--TTTTGCCATGTAATTACCTAAATTAGGTAATTACATGGCAAA
9ant	A,B	Antennapedia homeodomain	*D. melanogaster*	2.4	--TTTTGCCATGTAATTACCTAAATTAGGTAATTACATGGCAAA
6pax	A	Paired domain	*H. sapiens*	2.5	--------GATGACCTAATAGGGAAAATAACACGAA--------
1akh	A,B	Mat a-1/α-2	*S. cerevisiae*	2.5	----------TACATGTAAAAATTACATCATATGA---------
1b72	A,B	Homeodomain	*H. sapiens*	2.35	------ACTCTATGATTGATCGTAATGCGCAAACG---------
1b8l	A,B	Homeodomain	*D. melanogaster*	2.4	---------TTTACGTTTAAAAGCTTAATAAACG----------
1mnm	C,D	Mat α-2	*S. cerevisiae*	2.25	--------GATTACCTAATAGGGAAATTTACACG----------
					* * *
3. LacI repressor family					
1wet*†	A,C	Purine repressor	*E. coli*	2.6	AACGAAAACGTTTTCGT
1bdh†	A,C	Purine repressor	*E. coli*	2.7	TACGCAAACGTTTGCGT
1bdi†	A,C	Purine repressor	*E. coli*	3.0	TACGCAAACGTTTGCGT
1pnr†	A,C	Purine repressor	*E. coli*	2.7	AACGAAAACGTTTTCG-
2pua†	A,C	Purine repressor	*E. coli*	2.9	TACGCAAACGTTTGCGT
2pub†	A,C	Purine repressor	*E. coli*	2.7	TACGCAAACGTTTGCGT
2puc†	A,C	Purine repressor	*E. coli*	2.6	TACGCAAACGTTTGCGT
2pud†	A,C	Purine repressor	*E. coli*	2.6	TACGCAAACGTTTGCGT
2pue†	A,C	Purine repressor	*E. coli*	2.7	TACGCAAACGTTTGCGT
2puf†	A,C	Purine repressor	*E. coli*	3.0	TACGCAAACGTTTGCGT
2pug†	A,C	Purine repressor	*E. coli*	2.7	AACGAAAATTTTTCGT-
1vpw†	A,C	Purine repressor	*E. coli*	2.7	TACGCAAACGTTTGCGT
1qpz†	A,C	Purine repressor	*E. coli*	2.5	TACGCAAATTTTTCGT-
1zay†	A,C	Purine repressor	*E. coli*	2.7	TACGCAAACGTTTGCGT
					*** *** ***
4. Endonuclease FokI family‡					
1fok*	A	Endonuclease FokI	*F. okeanokoites*	2.8	TCGGATGATAACGCTAGTCA
5. γδ-resolvase family‡					
1gdt*	A,B	γδ-resolvase	*E. coli*	3.0	GCAGTGTCCGATAATTTATAAA
6. Hin recombinase family‡					
1hcr*	A	Artificial Hin recombinase	Artificial	1.8	TGTTTTTGATAAGA
7. RAP1 family					
1ign*	A,B	RAP1	*S. cerevisiae*	2.25	CCGCACACCCACACACCAG
8. Prd paired domain family					
1pdn*	C	Prd paired domain	*D. melanogaster*	2.5	AACGTCACGGTTGAC

(*Table 2 continues overleaf*)

Table 2 *(continued)*

PDB code	DNA-binding subunits	Protein name	Species	Resolution (Å)	DNA sequence
9. Tc3 transposase family‡					
1tc3*	C	Transposase TC3	*C. elegans*	2.45	AGGGGGGGTCCTATAGAACTT
10. Trp repressor family					
1trr*	A,B,D,E, G,H,J,K	Trp repressor	*E. coli*	2.4	AGCGTACTAGTACGCT
11. Diptheria Tox repressor family					
1ddn*	A,B,C,D	Diptheria tox repressor	*E. coli*	3.0	ATATAATTAGGATAGCTTTACCTAATTATTTTA
12. Transcription factor IIB					
1d3u*	B	TF IIB	*P. woesei*	2.4	AGAGTAAAGTTTAAATACTTATAT
1vol	A	TF IIB	*H. sapiens*	2.7	------GGCTATAAAAGGGCTG--
1c9b	A,E,I,M,Q	TF IIB	*H. sapiens*	2.65	--GGGCGCCTATAAAAGGGG----
					* * ****
'Winged' HTH					
13. Interferon regulatory factor					
2irf*	G,H,I, J,K,L	Interferon regulatory factor-2	*M. musculus*	2.2	AAGTGAAAGCGA
1if1	A,B	Interferon regulatory factor	*M. musculus*	3.0	AACTGTAAGCTTT
14. Catabolite gene activator protein family					
2cgp*	C	Catabolite gene activator	*E. coli*	2.2	--------------------------GTCACATTAAT----------
1ber	A,B	Catabolite gene activator	*E. coli*	2.5	--------GCGAAAAGTGTGACATATGTCACACTTTTCGGCGAAAAGTGTGACATATGTCACACTT
1cgp	A,B	Catabolite gene activator	*E. coli*	3.0	ACTTTTCGGCGAAAAGTGTGACATATGTCACACTTTTCGGCGAAAAGTGTGACATAT---------
1run	A,B	Catabolite gene activator	*E. coli*	2.7	--------GCGAAAAATGTGATCTAGATCACATTTTTCGGCGAAAAATGTGATCTAGATCACATTT
1ruo	A,B	Catabolite gene activator	*E. coli*	2.7	--------GCGAAAAATGTGATCTAGATCACATTTTTCGGCGAAAAATGTGATCTAGATCACATTT
					******* ***** ** ***** * ********** ***** *
15. Transcription factor family					
3hts*†	B	Heat shock transcription factor	*K. lactis*	1.75	GGTTCTAGAACC----
1cf7	A,B	E2F-DP	*H. sapiens*	2.6	-ATTTTCGCGCGGTTT
					** * * *
16. Ets domain family					
1bc8*	C	Sap-1	*H. sapiens*	1.93	TACCGGAAGTAACTTCCGGT
1bc7	C	Sap-1	*H. sapiens*	2.01	TACCGGAAGTAACTTCCGGT
1pue	E,F	TF PU.1	*M. musculus*	2.1	-AAAAAGGGGAAGTGGG---
1awc	A,B	GABP-α	*M. musculus*	2.15	-AACGACCGGAAGTACACCGGA
					* * ** *
II. Zinc-coordinating group					
17. ββα-zinc finger family					
1aay*	A1(103-133), A2(134-161), A3(162-187)	Zif268 zinc finger	*M. musculus*	1.6	--AGCGTGGGCGTTACGCC-CACGC----------
1zaa	C1(3-33), C2(34-61), C3(62-87)	Zif268 zinc finger	*M. musculus*	2.1	--AGCGTGGGCGTTACGCC-CACGC----------
2drp	A1(103-135), A2(137-164), D1(102-135), D2(137-165)	Tramtrack protein	*D. melanogaster*	2.8	AATAAGGATAACGTCCGTCGGACGTTATCCTTATTA--
1ubd	C1(295-323), C2(324-351), C3(352-381), C4(382-408)	Zinc finger	*H. sapiens*	2.5	-AGGGTCTCCAATTTTGAAGCGCGCTTCAAAATGG---

(Table 2 continues overleaf)

Table 2 *(continued)*

PDB code	DNA-binding subunits	Protein name	Species	Resolution (Å)	DNA sequence
1mey	C1(1-31), C2(32-59), C3(60-84) F1(1-31), F2(32-59), F3(60-84), G	Consensus zinc finger	Artificial	2.2	---ATGAGGCAGAACTTAGTTCTGCCTCA---------
1a1g	A1(103-132), A2(133-159) A3(160186)	DSNR (Zif268 variant)	*M. musculus*	1.9	---AGCGTGGGCGTTACGCCCACGC
1a1h	"	QSGR (Zif268 variant)	*M. musculus*	1.6	---AGCGTGGGCGTTACGCCCACGC
1a1l	"	RADR (Zif268 variant)	*M. musculus*	1.9	---AGCGTGGGCGTTACGCCCACGC
1a1j	"	"	*M. musculus*	1.6	---AGCGTGGGCGTTACGCCCACGC
1a1k	"	"	*M. musculus*	2.0	---AGCGTGGGCGTTACGCCCACGC
1a1l	"	Zif268 zinc finger	*M. musculus*	2.3	---AGCGTGGGCGTTACGCCCACGC
2gli	A1(103-135)	Five-zinc finger	*H. sapiens*	2.6	---AGCGTGGGCGTTACGCCCACGC
	A2(136-167)	"	"	"	---TTTCGTCTTGGGTGGTCCACG
	A3(168-196)	"	"	"	
	A4(197-228)	"	"	"	
	A5(229-257)	"	"	"	
					*
18. Hormone-nuclear receptor family					
2nll*	A,B	Retinoic acid receptor	*H. sapiens*	1.9	---CAGGTCAT-TTCAGGTCAGCTGACCTGAAATGACCTG--
1hcq	A,B,E,F	Estrogen receptor	*H. sapiens*	2.4	--CCAGGTCAC-AGTGACCTGCCAGGTCACTG-TGACCTG--
1glu	A,B	Glucococorticoid receptor	*R. norvegicus*	2.9	--CCAGAACATCGATGTTCTGCCAGAACATCGATGTTCTG--
1lat	A,B	Glucococorticoid receptor	*R. norvegicus*	1.9	TTCCAGAACATGTTCTGGATTCCAGAACAT----GTTCTGGA
1by4	A,B,C,D	Retinoic acid receptor	*H. sapiens*	2.1	---CAGGACATCTAGTAAATTCCAGATCTTACGTTGTCTG--
1cit	A	Orphan nuclear receptor	*H. sapiens*	2.7	--CCAGAACATCGAGCCTCTGCCAGAACATCGTTGTTCTG--
1a6y	A,B	Orphan nuclear receptor	*H. sapiens*	2.3	-TCCAGGACATCGCTAAGCTTGCTGGTCATTGCGGTTCTG
					*** ** * * * ***
19. Loop-sheet-helix family					
1tsr*	A,B,C	p53 tumor suppressor	*H. sapiens*	2.2	TTTCCTAGACTTGCCCAATTAATAATTGGGCAAGTCTAGGAA
1tup	A,B,C	p53 tumor suppressor	*H. sapiens*	2.2	TTTCCTAGACTTGCCCAATTAATAATTGGGCAAGTCTAGGAA
20. GAL4-type family:					
1zme*	C,D	Proline utilization	*S. cerevisiae*	2.5	-ACGGGAAGCCAACTCCG-
1d66	A,B	GAL4	*S. cerevisiae*	2.7	CCGGAGGACAGTCCTCCGG
					* * * *****
III. Zipper-type group					
21. Leucine zipper family					
2dgc*†	A,C	GCN4	*S. cerevisiae*	2.2	---TGGAGATGACGTCATCTCC--
1dgc†	A,C	GCN4	*S. cerevisiae*	3.0	---TGGAGATGACGTCATCTCC--
1ysa	C,D	GCN4	*S. cerevisiae*	2.9	---TTCCTATGAGCTCATCCAGTT
1a02	F	C-Fos	*H. sapiens*	2.7	TTGGGAAATTTCTTTCATAG----
	J	C-Jun			
					* ****
22. Helix-loop-helix family					
1am9*	A,B,C,D	Srebp-1A	*H. sapiens*	2.3	-TTGCAAGTGGGGTGATCCATGA---------------
1hlo	A,B	Max	*H. sapiens*	2.8	-CACCACGTGGTGTGGTGCACCA---------------
1an4	A,B	USF	*H. sapiens*	2.9	AGGCCACGTGACCGG-GGTACATCCGGTGCACT-----
1an2	A,C	Max	*M. musculus*	2.9	AGGTCACGTGACCTACACCACATCCAGTGCACTGGATG
1mdy	A,B,C,D	Myod	*M. musculus*	3.0	TCAACAGCTGTTGA---TCAACAGCTGTTGAC------
1a0a	A,B	Pho4	*S. cerevisiae*	2.8	
					** ** * ** **

(*Table 2 continues overleaf*)

Table 2 *(continued)*

PDB code	DNA-binding subunits	Protein name	Species	Resolution (Å)	DNA sequence
IV. Other α-helix group					
23. Papillomavirus-1 E2 family					
2bop*†	A,C	Papillomavirus-1 E2	Bovine papillomavirus	1.7	CCGACCGACGTCGGTCG
24. Histone family					
1aoi*	A,B,C,D, E,F,G,H	Histone	*X. laevis*	2.8	ATCAATATCCACCTGCAGATTCTACCAAAAGTGTATTTGGAAACTGCTC CATCAAAAGGCATGTTCAGCTGAATTCAGCTGAACATGCCTTTTGATGG AGCAGTTTCCAAATACACTTTTGGTAGAATCTGCAGGTGGATATTGAT
25. EBNA1 nuclear protein family					
1b3t*	A,B	Ebna-1	*H. herpesvirus 4*	2.2	GGGAAGCATAGCTTCCC
26. Skn-1 transcription factor					
1skn*	P	Skn-1	*C. elegans*	2.5	TGACAATGTCATCCC
27. Cre recombinase family‡					
1crx*	A,B	Cre recombinase	Bacteriophage P1	2.4	TATAACTTCGTATAG
28. High mobility group family					
1qrv*	A,B	High mobility group-1	*D. melanogaster*	2.2	GCGATATCGC------
1ckt	A	High mobility group-1	*R. norvegicus*	2.5	CCCCTCTGGACCTTCC * *
29. MADS box family					
1mnm*	A,B	Pheromone transcription factor MCM-1	*S. cerevisiae*	2.25	GATTACCTAATAGGGAAATTTACACG
V. β-sheet group					
30. TATA box-binding family					
1ytb*	B	TATA box-binding protein	*S. cerevisiae*	1.8	--GTATATAAAACGGGTGGCGTTTTATATAC-----
1ytf	A	TATA box-binding protein	*S. cerevisiae*	2.5	----TGTATGTATATAAAACGTTTTATATACATACA
1ais	A	TATA box-binding protein	*P. woesei*	2.1	AACTTACTTTIIAAAGCTTGAATGAAAAATTTCA--
1cdw	A	TATA box-binding protein	*H. sapiens*	1.9	CTGCTATAAAAGGCTGCAGCCTTTTATAGCAG----
1tgh	A	TATA box-binding protein	*H. sapiens*	2.0	--------CGTATATATACGCGTATATATACG----
1vol	B	TATA-box-BP	*H. sapiens*	2.7	----TGATCCCTTAAACTCGCTTGTATATGA-----
1d3u	A	TATA-box-BP	*P. woesei*	2.4	---CTTAATCGCTATATCCGTTTCTATAGCTTTCA-
1c9b	B,F,J,N,R	TATA-box-BP	*H. sapiens*	2.65	--GCTATAACGGTTAACGTTATTGTATAGCCAA--- * ****
VI. β-hairpin/ribbon group					
31. MetJ repressor protein					
1cma*	A,B	Met repressor	*E. coli*	2.8	TTAGACGTCTAGACGTCTA
32. Tus replication terminator family					
1ecr*	A	Tus replication terminator	*E. coli*	2.7	TTAGTTACAACATACT
33. Integration host factor family					
1ihf*	A,B	Integration host factor	*E. coli*	2.5	CGGTGCAACAAATTGATAAGCAATGCTTTTTTGGC
34. Transcription factor T-domain					
1xbr*	A,B	T-domain	*X. laevis*	2.5	AATTTCACACCTAGGTGTGAAAATT
35. Hyperthermophile DNA-BP.					
1azp*	A	Sac7D	*S. acidocaldarius*	1.6	GCGATCGC
1azq	A	Sac7D	*S. acidocaldarius*	1.94	GCGATCGC
1bnz	A	7A	*S. acidocaldarius*	2.0	GTAATTAC
1bf4	A	Ss07D	*S. acidocaldarius*	1.6	GTAATTAC * ** *

(*Table 2 continues overleaf*)

Table 2 *(continued)*

PDB code	DNA-binding subunits	Protein name	Species	Resolution (Å)	DNA sequence
36. Arc repressor					
1bdt*	A,B,C,D	Arc repressor	Bacteriophage P22	2.5	TATAGTAGAGTGCTTCTATCAT
1bdv	A,B,C,D	Arc repressor	Bacteriophage P22	2.8	TATAGTAGAGTGCTTCTATCAT
1par	A,B,C,D	Arc repressor	Bacteriophage P22	2.6	TATAGTAGAGTGCTTCTATCAT
VII. Other					
37. Rel homology region family					
1a3q*	A, B	NF κB p52	*H. sapiens*	2.1	GATTCCCCGGGGAATTCCCC------
1nfk	A,B	NF-κB p50	*M. musculus*	2.3	GAATTCCCTGGGAATTCCC-------
1svc	P,T	NF-κB p50	*H. sapiens*	2.6	----AGATGGGGAATCCCCTAGA---
1vkx	A,B	NF κB p50/p65	*M. musculus*	2.9	------CTGGGGAATTTCCCAG----
1ram	A,B	NF κB p65	*M. musculus*	2.7	GGGACTTTCCGAAATTCCCC------
1bvo	A	GambifI TF	*A. gambiae*	2.7	---CGGGCTGGAAATTTCCAGCCG--
1a02	N	Nfat	*H. sapiens*	2.7	------TTGGGAAATTTCTTTCATAG * *** *
38. Stat protein family					
1bf5*	A	Stat-I	*H. sapiens*	2.9	ACAGTTTCCCGTAAATGC
VIII. Enzyme group					
39. Methyltransferase family					
6mht*	A	HhaI methyltransferase	*H. haemolyticus*	2.05	---GATAGCGCTATCTGATAGCGCTATC----------
4mht	A	HhaI methyltransferase	*H. haemolyticus*	2.7	---GATAG-GCTATC-GATAGCGCTATC----------
1mht	A	HhaI methyltransferase	*H. haemolyticus*	2.8	---GATAGCGCTATCTGATAGCGCTATC----------
3mht	A	HhaI methyltransferase	*H. haemolyticus*	2.7	---GATAGCGCTATCTGATAGCGCTATC----------
5mht	A	HhaI methyltransferase	*H. haemolyticus*	2.7	---GATAGCGCTATCTGATAGCGCTATC----------
7mht	A	HhaI methyltransferase	*H. haemolyticus*	2.87	---GATAGCGCTATCTGATAGCGCTATC----------
8mht	A	HhaI methyltransferase	*H. haemolyticus*	2.76	---GATAGCGCTATCTGATAGCGCTATC----------
9mht	A	HhaI methyltransferase	*H. haemolyticus*	2.87	---GATAG-GCTATC-GATAGCGCTATC----------
10mh	A	HhaI methyltransferase	*H. haemolyticus*	2.55	---GATAGCGCTATCTGATAGCGCTATC----------
1dct	A,B	HaeIII methyltransferase	*H. aegyptus*	2.8	ACCAGCAG-GCCACC-AGTGTCACTGGTGGCCTGCTGG ** ** * * * * **
40. Endonuclease PvuII family					
3pvi*	A,B	Endonuclease PvuII	*P. vulgaris*	1.59	TGACCAGCTGGTC
2pvi	A,B	Endonuclease PvuII	*P. vulgaris*	1.76	TGACCAGCTGGTC
1pvi	A,B	Endonuclease PvuII	*P. vulgaris*	2.8	TGACCAGCTGGTC
41. Endonuclease EcoRV family					
1rva*	A,B	Endonuclease EcoRV	*E. coli*	2.0	------AAAGATATCTTAAA---GATATCTT-
1rvb	A,B	Endonuclease EcoRV	*E. coli*	2.1	------AAAGATATCTTAAA---GATATCTT-
1rvc	A,B	Endonuclease EcoRV	*E. coli*	2.1	------AAAGATATCTTAAA---GATATCTT-
2rve	A,B	Endonuclease EcoRV	*E. coli*	3.0	CGAGCTCGCGAGCTCGCGAGCTCGCGAGCTCG
4rve†	A,B,C,G	Endonuclease EcoRV	*E. coli*	3.0	-------GGGATATCCCGG----GATATCCC-
1rve	A,B	Endonuclease EcoRV	*E. coli*	2.5	------AAAGATATCTTAAA---GATATCTT-
1rv5	A,B	Endonuclease EcoRV	*E. coli*	2.1	------CGGGATATCCC---CGGGATATCCC-
1bgb	A,B	Endonuclease EcoRV	*E. coli*	2.0	------AAAGATATCTTAAA---GATATCTT-
1bss	A,B	Endonuclease EcoRV	*E. coli*	2.0	------AAAGATATCTTAAA---GATATCTT-
1bua	A,B	Endonuclease EcoRV	*E. coli*	2.15	------AAAGACATCTT---------------
1bsu	A,B	Endonuclease EcoRV	*E. coli*	2.0	------AAAGACATCTT--------------- ** ** * * *
42. Endonuclease EcoRI family					
1qps*†	A,M	Endonuclease EcoRI	*E. coli*	2.5	TCGCGAATTCGCG
1eri†	A,C	Endonuclease EcoRI	*E. coli*	2.7	TCGCGAATTCGCG
1qrh†	A	Endonuclease EcoRI	*E. coli*	2.5	TCGCGAATTCGCG
1qri†	A	Endonuclease EcoRI	*E. coli*	2.6	TCGCGAATTCGCG
43. Endonuclease BamHI family					
3bam*	A,B	Endonuclease BamHI	*B. amyloliquefaciens*	1.8	TATGGATCCATATATG
1bhm	A,B	Endonuclease BamHI	*B. amyloliquefaciens*	2.2	TATGGATCCATA---- ************

(*Table 2 continues overleaf*)

Table 2 *(continued)*

PDB code	DNA-binding subunits	Protein name	Species	Resolution (Å)	DNA sequence
44. Endonuclease V family					
1vas*	A	Endonuclease V	Phage T4	2.75	ATCGCGTTGCGCTTAGCGCAACGCCGA
45. Dnase I family					
2dnj*	A	Deoxyribonuclease I	*B. taurus*	2.0	GCGATCGCGCGATC--
1dnk	A	Deoxyribonuclease I	*B. taurus*	2.3	GGTATACGGCTATACC * ** * **
46. DNA mismatch endonuclease					
1cw0*	A	DNA mismatch endonuclease	*E. coli*	2.3	ACGTACCTGGCTAGCTAGGTACGT
47. DNA polymerase- β family					
1bpy*	A	DNA polymerase-β	*H. sapiens*	2.2	CCGACCACGCATCAGC
1bpx	A	”	”	2.4	CCGACCACGCATCAGC
1bpz	A	”	”	2.6	CCGACCACGCATCAGC
1zqa	A	”	”	2.7	CATTAGAATCTAATG-
1zqf	A	”	”	2.9	CATTAGAATCTAATG-
1zqi	A	”	”	2.7	CATTAGAATCTAATG-
1zqn	A	”	”	3.0	CATTAGAATCTAATG-
1zqp	A	”	”	2.8	CATCTG--TCAGATG-
7ice	A	”	”	2.8	CATCTG--TCAGATG-
7icg	A	”	”	3.0	CATCTG--TCAGATG-
7ich	A	”	”	2.9	CATCTG--TCAGATG-
7ici	A	”	”	2.8	CATCTG--TCAGATG-
7ick	A	”	”	2.9	CATCTG--TCAGATG-
7icm	A	”	”	3.0	CATCTG--TCAGATG-
7icn	A	”	”	2.8	CATCTG--TCAGATG-
7icp	A	”	”	3.0	CATCTG--TCAGATG-
7icq	A	”	”	2.9	CATCTG--TCAGATG-
7icr	A	”	”	3.0	CATCTG--TCAGATG-
7ics	A	”	”	2.8	CATCTG--TCAGATG-
7ict	A	”	”	2.8	CATCTG--TCAGATG-
7icv	A	”	”	2.8	CATCTG---CAGATG-
8ica	A	”	”	3.0	CATTAGAATCTAATG-
8icc	A	”	”	2.8	CATTAGAATCTAATGA
8icf	A	”	”	2.9	CATTAGAATCTAATG-
8ici	A	”	”	2.8	CATTAGAATCTAATGA
8ick	A	”	”	2.7	CATTAGAATCTAATG-
8icm	A	”	”	2.9	CATTAGAATCTAATGA
8icn	A	”	”	2.8	CATTAGAATCTAATG-
8ico	A	”	”	2.7	CATTAGAATCTAATGA
8icp	A	”	”	2.9	CATTAGAATCTAATG-
8icq	A	”	”	3.0	CATTAGAATCTAATGA
8icr	A	”	”	2.9	CATTAGAATCTAATGA
8ics	A	”	”	2.9	CATTAGAATCTAATGA
8icu	A	”	”	3.0	CATTAGAATCTAATG-
8icx	A	”	”	3.0	CATTAGAATCTAATG-
9ica	A	”	”	3.0	CATTAGAA-CTAATG-
9icf	A	”	”	3.0	CATTAGAT-CTAATG-
9icg	A	”	”	3.0	CATTAGAT-CTAATG-
9ich	A	”	”	2.9	CATTAGAT-CTAATG-
9ick	A	”	”	2.7	CATTAGAT-CTAATG-
9icl	A	”	”	2.8	CATTAGAT-CTAATG-
9icm	A	”	”	2.9	CATTAGAT-CTAATG-
9icn	A	”	”	3.0	CATTAGAATCTAATG-
9ico	A	”	”	2.9	CATTAGAATCTAATG-
9icq	A	”	”	2.9	CATTAGAATCTAATG-
9icr	A	”	”	3.0	CATTAGAA-CTAATG-
9ics	A	”	”	2.9	CATTAGAT-CTAATG-
9ict	A	”	”	3.0	CATTAGAT-CTAATG-
9icu	A	”	”	2.9	CATTAGAATCTAATG-
9icv	A	”	”	2.7	CATTAGAATCTAATG-
9icw	A	”	”	2.6	CATTAGAATCTAATG-
9icx	A	”	”	2.6	CATTAGAATCTAATG-
9icy	A	”	”	3.0	CATTAGAATCTAATG-
2bpf	A	”	*R. norvegicus*	2.9	GGGCGCCGCGGCGCC- * * *

(*Table 2 continues overleaf*)

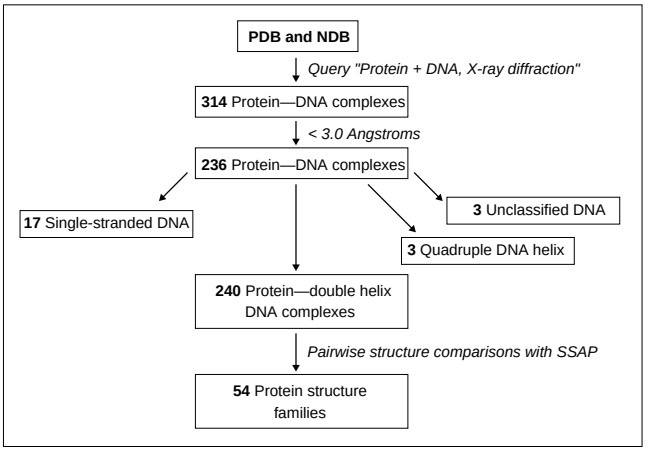
Table 2 *(continued)*

PDB code	DNA-binding subunits	Protein name	Species	Resolution (Å)	DNA sequence
48. DNA polymerase I					
2bdp*	A	DNA polymerase I	*B. stearothermophilus*	1.8	--GCATGATGCAGCATCATGC---
4bdp	A	DNA polymerase I	*B. stearothermophilus*	1.8	--GCATGATGCAGCATCATGC---
1qsy	A	DNA polymerase I	*T. aquaticus*	2.3	GACCACGGCGCAGGGCGCCGTGGT
1qss	A	DNA polymerase I	*T. aquaticus*	2.3	GACCACGGCGCAGGGCGCCGTGGT
2ktq	A	DNA polymerase I	*T. aquaticus*	2.3	GACCACGGCGCGGGCGCCGTAATC
3ktq	A	DNA polymerase I	*T. aquaticus*	2.3	GACCACGGCGCGGGCGCCGTAATC
4ktq	A	DNA polymerase I	*T. aquaticus*	2.5	GACCACGGCGCGGGCGCCGTAATC
1tau	A	DNA polymerase I	*T. aquaticus*	3.0	-GCCATGCGGCAGTCGC-------
1d8y	A	DNA polymerase I	*E. coli*	2.08	TTTTTTTTTTTTTTTTT ** * ** *
49. DNA polymerase T7					
1t7p*	A	DNA polymerase	Bacteriophage	2.2	GCCAGTGCCAACCTTGGCACTGGC
1clq	A	DNA polymerase	Bacteriophage	2.7	GCGGAACTACTAGTAGTTCCGAG ** * * * *
50. HIV reverse transcriptase					
1hmi*	A,B	HIV reverse transcriptase	HIV type 1	2.8	ATGGCGCCCGAACAGGAC
2hmi	A,B	HIV reverse transcriptase	HIV type 1	2.8	ATGGCGCCCGAACAGGAC
51. Uracil-DNA glycosylase					
1ssp*	E	Uracil-DNA glycosylase	*H. sapiens*	1.9	CTGTATCTTAAAGATAACAG
2ssp	E	Uracil-DNA glycosylase	*H. sapiens*	2.25	CTGTATCTTAAAGATAACAG
4skn	A	Uracil-DNA glycosylase	*H. sapiens*	2.9	TGGGGGCTTAAAGCCGCCC- * ******* *
52. 3-Methyladenine DNA glycosylase					
1bnk*	A	3-Methyladenine DNA glycosylase H.	*H. sapiens*	2.7	GACATGTTGCCTGGCAATCATGTCA
53. Homing endonuclease					
1a73*	A,B,	Intron-encoded Endonuclease I-Ppoi	*P. polycephalum*	1.8	-TTGACTCTCTTAAGCGAGTCA
1a74	A,B,	Intron-encoded Endonuclease I-Ppoi	*P. polycephalum*	2.2	-TTGACTCTCTTAAGCGAGTCA
1cyq	A,B,	Intron-encoded Endonuclease I-Ppoi	*P. polycephalum*	1.93	-TTGACTCTCTTAAGCGAGTCA
1ipp	A,B,	Intron-encoded Endonuclease I-Ppoi	*P. polycephalum*	2.2	-TTGACTCTCTTAAGCGAGTCA
1bp7	A,B,C,D	Homing endlonuclease I-CreI	*C. reinhardtii*	3.0	GCAAAACGTCGTGAGACAGTTCG * ** * ** ***
54. Topoisomerase I					
1a31*	A	Topoisomerase I	*H. sapiens*	2.1	AAAAAGACCCTGAAAAACCCCT
1a35	A	Topoisomerase I	*H. sapiens*	2.5	AAAAAGACCCTGAAAAACCCCT
1a36	A	Topoisomerase I	*H. sapiens*	2.8	AAAAAGACTTAGAAAAATTTTT

Each entry is provided with the four-digit PDB code, the name, the source, resolution and the aligned sequences of DNA to which the protein is bound in the crystal structure. Protein DNA-binding domains are identified by a single letter identifying the protein chain in the PDB file. Where more than one DNA-binding domain is contained in a single subunit, the chain is split into the individual domains. For these, the domains are labeled with the chain ID and a number and the PDB file residue numbers of the domain are given in brackets. The PDB entries are classified into 'groups' and 'families' according to the structure of the protein contained in the complex. The representative for each family, defined as the structure with the highest resolution is marked with an asterisk (*). The CAP family is an exception; the highest-resolution entry, 2cgp, contains only part of the dimer and so PDB code 1ber is used as the representative. The full coordinate files for homodimeric PDB entries that contain only half the structure were obtained from the NDB and are marked by a dagger (†). Enzyme families whose structures also qualified for another protein group are listed in all that are applicable and are marked by a double dagger (‡).

Where a protein subunit is specified, the corresponding chain identity in the PDB file is added to the four-digit code (for example, 1aayA). For a particular segment within a subunit, an identifier number, as defined in Table 2, is added (for example, 1aayA1).

The result is a total of 54 protein families of which 33 contain more than one PDB entry. Within each family there are structures of the same protein bound to different DNA sequences (for example, the phage 434 repressor complexes 1per and 1rpe in the Cro and Repressor family) and structures of different proteins bound to different DNA sequences, (for example, the phage 434 and λ repressor complexes, 1per and 1lli respectively, in the Cro and Repressor family). Table 2 lists all the protein-DNA complex structures in the dataset and their classifications.



Box 1
Flow diagram showing the selection of the protein-DNA complexes from the PDB (04/01/00). The protein-DNA complexes were groups into structurally related families using the secondary structure alignment program SSAP (see text).

Also shown are multiple alignments of the DNA sequences that are bound in each family, as computed by ClustalX [11].

Here we review the eight groups of protein-DNA complexes listed above and their individual families.

Group I: helix-turn-helix proteins

The HTH motif is a common recognition element used by transcription regulators and enzymes of prokaryotes and eukaryotes [1,2,12-15]. Although the motif is traditionally defined as a 20-amino-acid segment of two almost perpendicular α helices connected by a four-residue β turn (Cro and Repressor family, 1lmb; Figure 1a) here we extend the definition to those with longer linkers, such as loops, as long as the relative orientation of the α helices is maintained (for example, the RAP1 protein family, 1ign). Examples from each family within the HTH group [16,17] are shown in Figure 1. In the figure, the HTH is highlighted in red.

The motif invariably binds in the DNA major groove; the second α helix, commonly known as the recognition, or probe, helix, is inserted in the groove. In most complexes, direct contacts are made between amino-acid side chains and nucleotide bases; in a few examples, however, protein backbone atoms or bridging water molecules are used (for example, the Trp repressor family, 1trrA). Supporting contacts with the DNA backbone are mainly made by the linker and the first α helix in the motif, which bridges the major groove at the amino-terminal end of the recognition helix. Further interactions with the nucleic acid can also be made by the rest of the protein and sometimes contribute to further specification of the DNA sequence. For example, the Hin recombinase protein (1hcrA) interacts with bases in the minor grooves adjacent to the one bound by the recognition helix [18].

The HTH motif is typically found in a bundle of three to six α helices, which provides a stabilizing hydrophobic core. Although motifs from different protein families are structurally very similar [19] little homology is observed outside the motif. In structures such as the 434 repressor protein (1lli, Cro and Repressor family), the HTH motif is part of the main body of the protein [20]. In others, such as the purine repressor (1wet, LacI repressor family), it is placed in a small domain extending out of the main structure [21]. There is little sequence similarity between the motifs of different families and this variation allows them to recognize distinct sets of DNA sequences.

The precise positioning of the recognition helix in the DNA major groove also varies, reflecting the structural and functional requirements of each protein. The recognition helices of prokaryotic transcription factors (for example, those of the Cro and Repressor family, such as 1lli) are generally aligned with their axes parallel to base-pairing edges of the nucleotides, whereas those of eukaryotic proteins (for example, the homeodomain family, see 1oct) are parallel to the sugar-phosphate backbone in order to accommodate the longer α helices [22]. Binding by the Trp repressor (1trrA) is unique, with the amino-terminal end of the α helix practically pointing into the groove. Although this last arrangement limits the role of amino acids further down the α helix, it is needed in order to allow a second repressor subunit into the same major groove when binding in tandem [23]. Helix binding in the major groove, which is also very common in other groups, provides a geometrically favorable framework in which components in both protein and DNA can change to allow multispecific complementarity. The protein sequences and the modes of interaction vary considerably, but the need for the helix to be 'presented' on the surface of the protein, ready for interaction with DNA, is satisfied by the HTH motif.

In general, the prokaryotic transcription factors bind to palindromic DNA sequences as homodimers, whereas eukaryotic proteins, such as members of the homeodomain family, bind both as monomers or heterodimers to non-symmetrical target sites. The latter arrangement potentially allows recognition of a much wider range of DNA sequences. The prokaryotic enzymes in the group (for example, *Fok*I endonuclease, 1fok) which function as monomers possess more than one motif in a single subunit.

There are 16 homologous families in the HTH group. Eight contain only one structure each, and, of the remaining six, only the Cro and Repressor and homeodomain families contain proteins with different amino-acid sequences. The

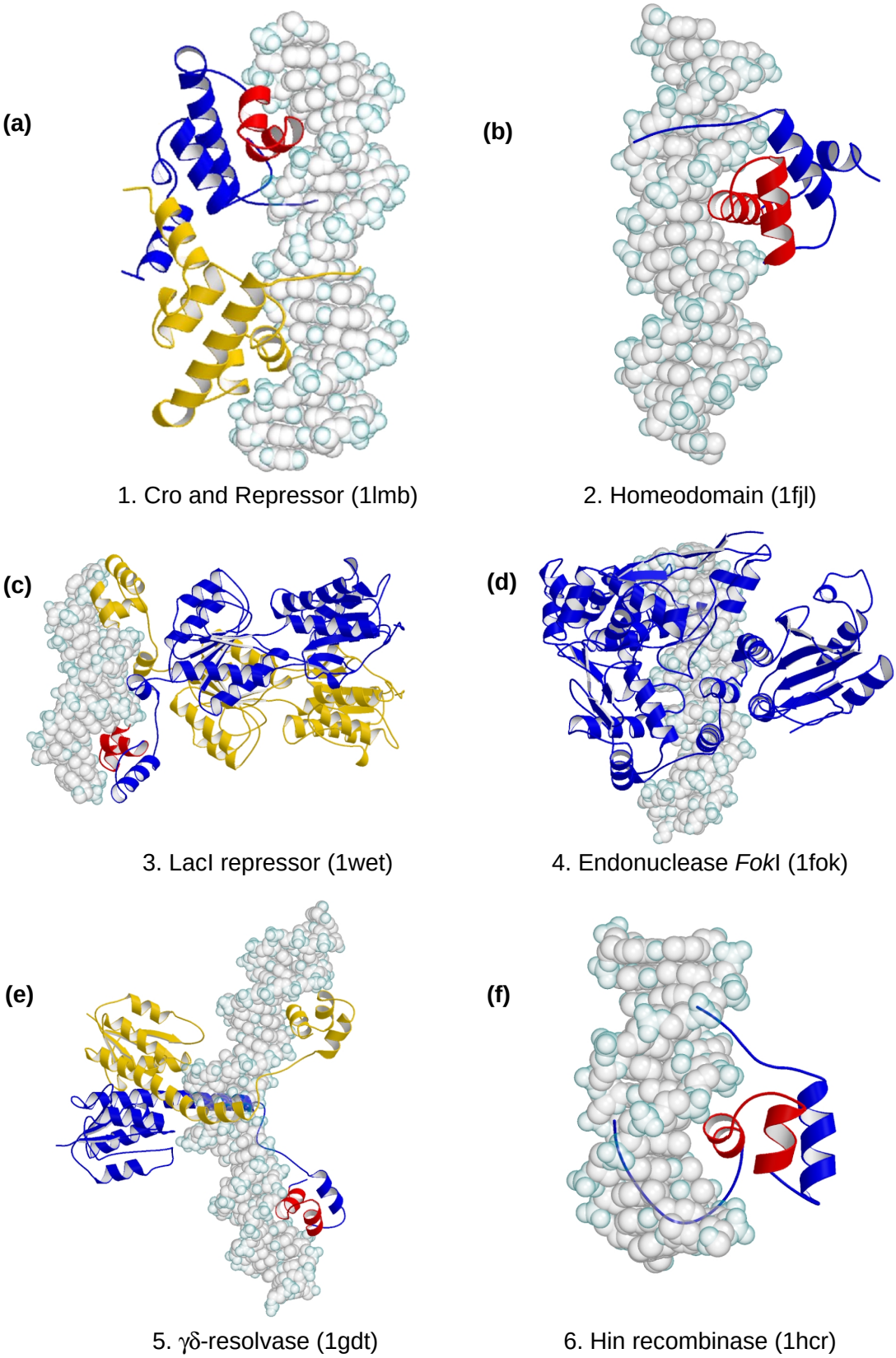
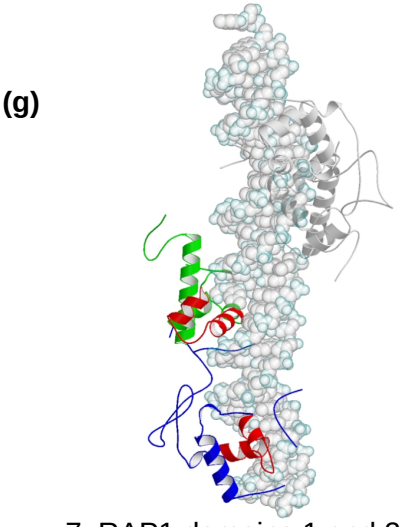
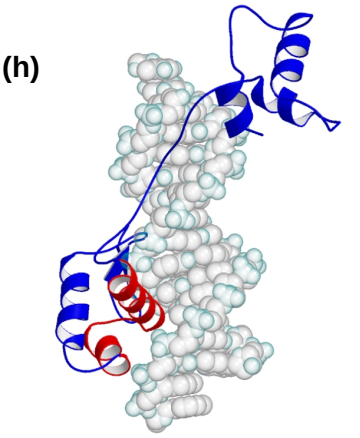


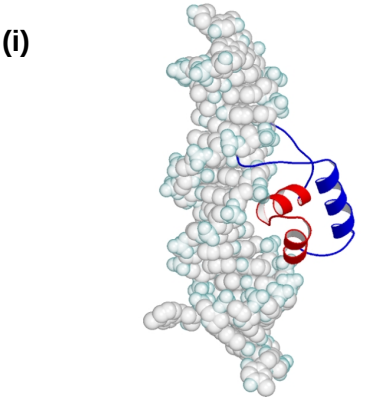
Figure 1 *(continues overleaf)*
Group 1, HTH proteins. The DNA-binding motif is red. The protein binds as a dimer; one monomer is colored blue and the other yellow. The DNA is shown as a space-filling model. Family names and numbers are as listed in Table 2; PDB codes are bracketed.



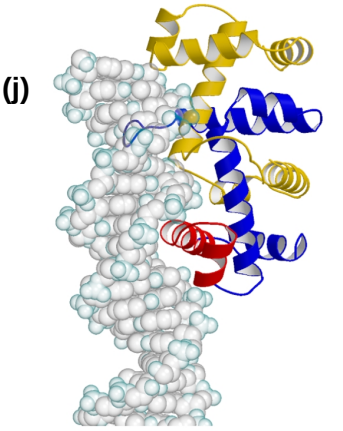
7. RAP1 domains 1 and 2 (1ign)



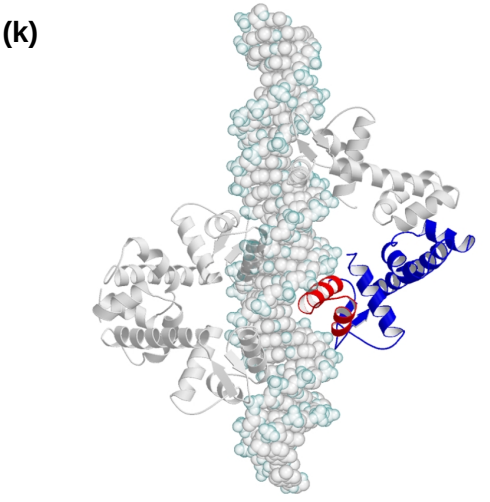
8. Prd paired domain (1pdn)



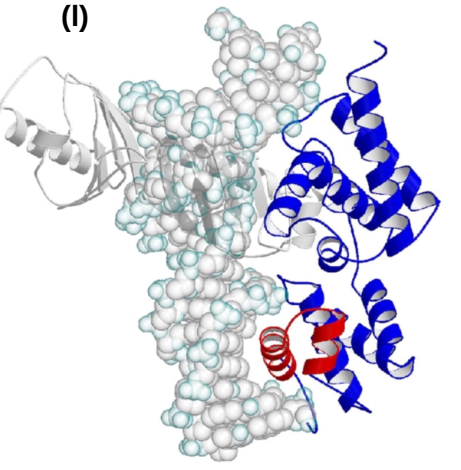
9. Tc3 transposase (1tc3)



10. Trp repressor (1trr)



11. Diptheria toxin repressor (1ddn)



12. Transcription factor IIB (1d3u)

Figure 1 *(continued)*
Group 1, HTH proteins.

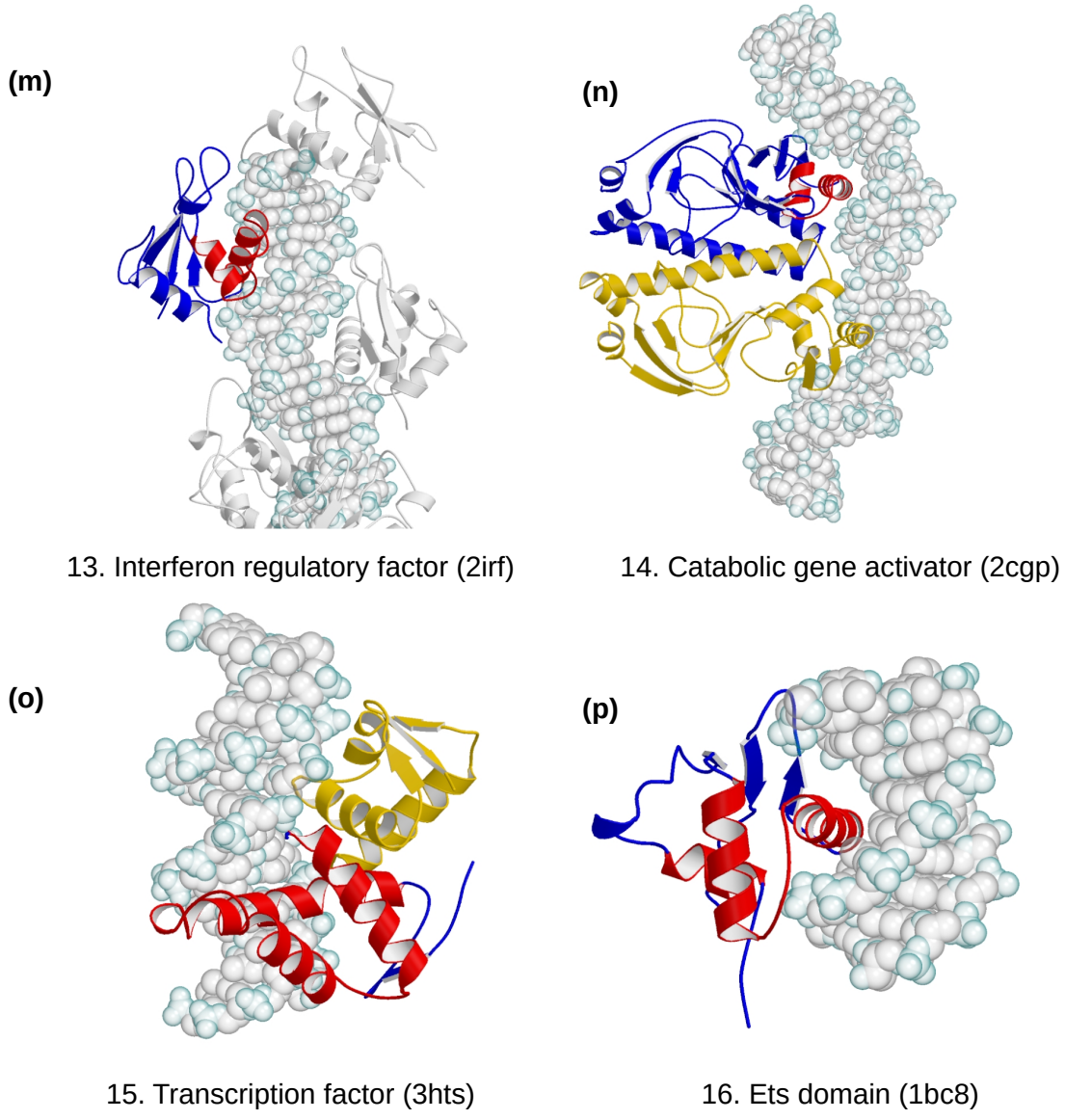


Figure 1 *(continued)*
Group 1, HTH proteins.

pairwise sequence identities between the subunits in the Cro and Repressor family range from 68% (1lmbA and 1perA) to 100% for identical proteins (1lliA and 1lmbA). Pairwise SSAP scores are above 85. In the homeodomain family, although POU domain proteins are often considered separately, they have been included together in this study because of the high SSAP scores that are found between the proteins. For example, the Matα-2 protein (1aplA) and the POU domain protein Pit-1 (1au7A1) have an SSAP score of 88.3 in an alignment of 59 protein residues. As a result, there is greater variation in pairwise sequence identities, which are as low as 42% (1aplA and 1au7A1). The Hin recombinase, γδ-resolvase, *Fok*I restriction endonuclease, Tc3 transposase and Cre recombinase families belong to both the HTH and enzyme groups.

'Winged' HTH proteins

The 'winged' HTH motif is an extension of the HTH group which is characterized by the presence of a third α helix and an adjacent β sheet (Figure 1m-p), which are considered to be components of the DNA-binding motif. The recognition helix binds as in the regular HTH motifs, and the extra secondary structural elements provide additional contacts with the DNA backbone.

Group II: zinc-coordinating proteins

Zinc-coordinating proteins make up the largest single group of transcription factors in eukaryotic genomes, and the DNA-binding motif is characterized by the tetrahedral coordination of one or two zinc ions by conserved cysteine and histidine

residues [1,2,15]. The widespread use of this arrangement is believed to be due to the structural stability the metal ions offer to domains that are not sufficiently large for a stable hydrophobic core [24]. The use of zinc-coordinating motifs is not limited to DNA binding and they are also found in domains that mediate protein-protein interactions [25]. Proteins in this group are more structurally diverse than those of the HTH group, and six principal families have been identified so far, of which four are represented in the dataset of complexes. The representative structures are shown in Figure 2, with the zinc-coordinating motif colored red. To avoid confusion over the use of the term 'zinc finger', we reserve its use for proteins that have the Zif-268-style (1aayA1) motif with two β strands and an α helix (Figure 2a). The name 'zinc-coordinating' will be used as the generic term for all proteins with zinc ions in the DNA-binding motif.

The ββα zinc-finger family

The ββα zinc-finger proteins constitute the largest individual family in the group and more than a thousand distinct sequence motifs have been identified in transcription factors [26]. The structure of the finger is characterized by a short two-stranded antiparallel β sheet followed by an α helix (Figure 2a). Two pairs of conserved histidine and cysteine residues in the α helix and second β strand coordinate a single zinc ion.

Protein subunits often contain multiple fingers that wrap round the DNA in a spiral manner. Fingers bind adjacent 3 bp subsites by inserting the α helix in the major groove, and the recognition pattern between the helix and DNA is well characterized. Amino acids at positions -1, 2, 3 and 6 relative to the start of the α helix are used to interact with the bases, -1 being the position that precedes the helix [27,28]. Although there are examples of complexes that do not follow this pattern [29], mutagenesis experiments have shown that by altering the amino acids at the key positions, different subsite sequences are recognized [30]. By adjusting the number of fingers in a protein, binding sites of varying lengths can be bound with different specificities. For example, a protein with five fingers is expected to bind a long target site very selectively, whereas a protein with only a single finger could potentially bind a wide range of sites containing the required subsite sequence. However, the structure of the human glioblastoma protein (1gli) suggests that binding is not always straightforward; of the five fingers in the structure, one does not contact the DNA at all, and only two appear to make specific contacts with bases [31].

As described earlier, the protein subunits in this study have been split into distinct domains, each containing a single zinc-finger motif. The pairwise sequence identities of the aligned domains are all high, ranging from 73% (for example, human zinc-finger protein, 1udbA1, and *Drosophila* tramtrack protein, 2drpA1) to 100% (for example, mouse Zif268 protein, 1aayA1, and artificial protein, 1mey). All domains are structurally very similar, returning SSAP scores of over 90.

Hormone receptor family

Nuclear receptors for steroid hormones, thyroid hormones and retinoids form the second family in the group (Figure 2b). On binding the appropriate ligand, these receptors translocate from the cytoplasm to the nucleus and regulate transcription at DNA sequences called hormone response elements [2,32]. Hormone receptors normally function as homo- or hetero-dimers and each monomer typically consists of ligand-binding, DNA-binding, and transcription regulatory domains. The zinc-coordinating motif is found in the DNA-binding domain and is characterized by two antiparallel α helices capped by loops at their amino-terminal ends; each helix-loop pair coordinates a single zinc ion using four conserved cysteines. The two α helices lie approximately at right angles to each other; the first is inserted in the DNA major groove to provide interactions with bases, while the loops and the second α helix contact the DNA backbone. The DNA-binding domain alone is sufficient for dimerization, and the interface is provided by the loops leading into the second α helix.

All receptor subunits recognize one of two half-site sequences, 5′-AGAACA-3′ or 5′-AGGTCA-3′. The identity of the full target site is determined by the two half-site sequences that are present, their relative orientation (either symmetric or palindromic) and the spacing between them (between 3 and 6 bp). Thus recognition of the target sequence depends on the read-out of the half-site sequences by each subunit and the manner in which the two subunits dimerize [33]. The sequences of all entries in the current dataset are very similar (sequence identities > 90%) except for the thyroid hormone receptor (for example, 1bsx), which has two extra helices in the carboxy-terminal tail. The structures are all very similar, with pairwise SSAP scores of over 90.

Loop-sheet-helix family

The third family of zinc-coordinating motifs is the loop-sheet-helix zinc-coordinating motif (Figure 2c). This is represented by the DNA-binding region of the protein p53, a transcriptional activator implicated in tumor suppression [2,34]. As the name indicates, the DNA-binding domain consists of a loop leading out of the main body of the protein, followed by a small β sheet, an α helix and then another loop that leads back into the protein. Three cysteines and a histidine in the two loop regions coordinate the zinc ion.

The protein binds with the α helix in the DNA major groove and the loops in the minor groove, although the latter are not thought to confer specificity. The protein functions as a tetramer with each subunit contacting a

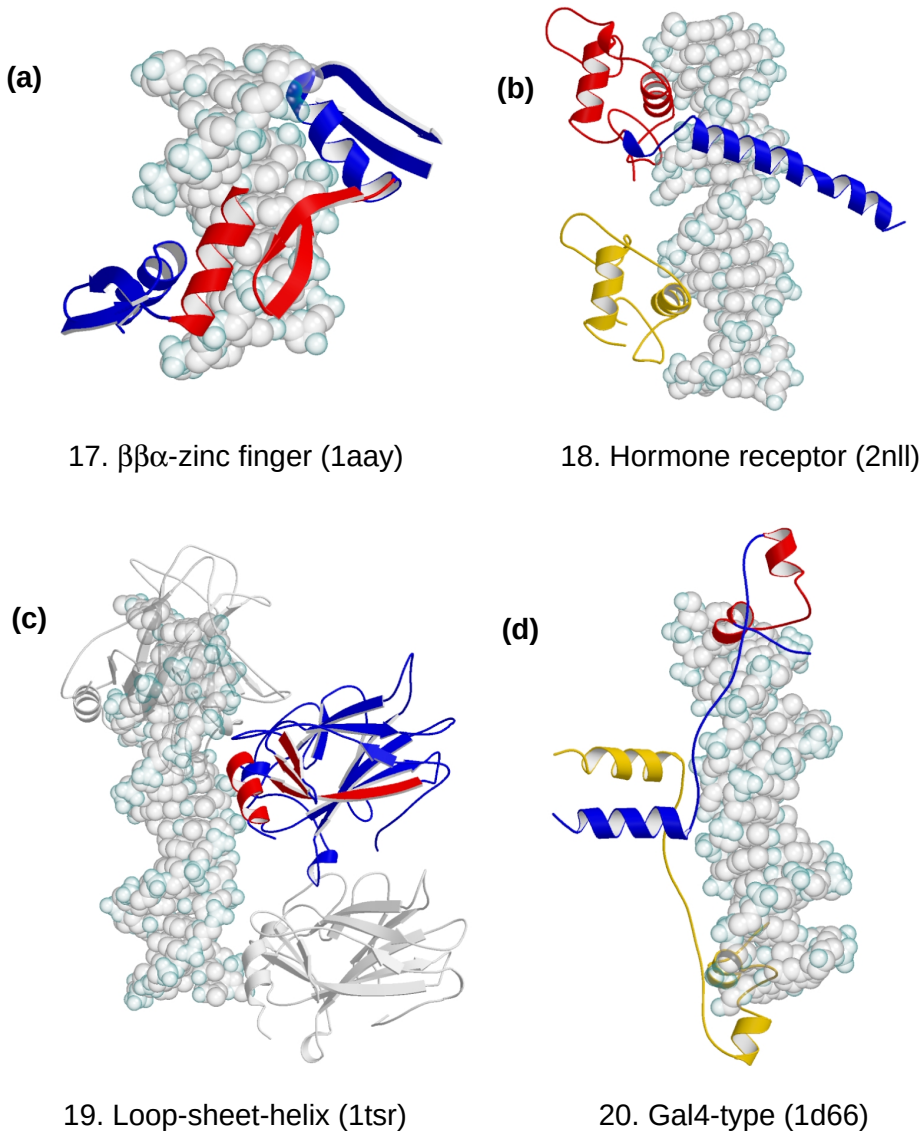


Figure 2
Group II, zinc-coordinating proteins. Colors, numbers and names are used as in Figure 1.

separate 5 bp recognition sequence positioned one after another. Regions outside the DNA-binding motif make the intersubunit interactions.

Gal4 family

The final zinc-coordinating family contains only the Gal4 protein [35]. It is a transcriptional regulator of galactose-induced genes and its zinc-coordinating motif has so far only been identified in yeast proteins. The motif comprises a pair of α helices that coordinate two zinc ions through six cysteine residues, where two of the cysteines are shared by both metal atoms (Figure 2d). The first α helix is presented in the DNA major groove for binding with bases, and the second α helix makes the backbone interactions. Gal4 functions as a homodimer, and the dimerization interface is located outside the zinc-coordinating motif.

Group III: zipper-type proteins

The zipper-type group derives its name from the method of dimerization used by its members, which so far only comprise those from eukaryotic organisms. Two families, the leucine zipper (Figure 3a) and helix-loop-helix proteins (Figure 3b), are defined; the latter must not be confused with the HTH group described earlier. While some members are reported to function as heterodimers (for example the Fos-Jun complex), all the PDB entries in the current dataset are of homodimers.

Leucine zipper family

In the leucine zipper family, the structure of the protein can be split into two parts: the dimerization region and the DNA-binding region. As shown in Figure 3a, each subunit in the leucine zipper protein consists of a single α helix about 60 amino acids long. Dimerization is mediated through the formation of a coiled coil by a 30-amino-acid section at the carboxy-terminal end of each helix. The segment, known as the zipper region, consists of leucine or a similar hydrophobic amino acid every eight residue positions - roughly every two turns of the α helix. Corresponding side chains from each subunit mediate hydrophobic contacts at the interface through side-by-side packing. The DNA-binding region, also known as the basic region, is found in the amino terminus and for the leucine zipper proteins, the binding segment is a direct extension of the dimerization region. The α helices of the two subunits diverge from the coiled coil and enter the DNA major groove in opposing directions, each binding to half of the target [36]. The leucine zipper family consists entirely of the yeast GCN4 proteins, which have near-identical structures and bind promoter regions of genes that encode enzymes involved in amino-acid biosynthesis.

Helix-loop-helix family

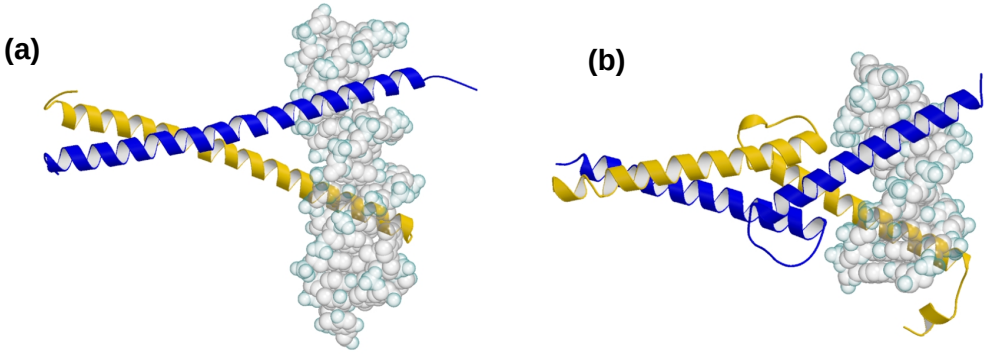
As the name suggests, helix-loop-helix proteins are a modification of the continuous α helices of the leucine zipper proteins in which the DNA-binding and dimerization regions are separated by a loop, resulting in a four-helix bundle (Figure 3b). Like those of leucine zippers, the dimerization helices interact with each other in a coiled-coil arrangement and the DNA-binding helices are inserted into the DNA major groove. By separating the two segments, more flexibility is allowed in positioning the probe helices on the nucleic acid [37,38]. The helix-loop-helix family is represented by the mouse and human forms of Max, mouse MyoD, and human USF proteins. Sequence identities range from 66% (Max protein, 1an2A, and USF protein, 1an4A) to 97% (mouse Max protein, 1an2A, and human Max protein, 1hloA) and with the exception of the MyoD (1mdyA) and USF (1an4A) protein pair (pairwise SSAP score 70), SSAP scores are above 80. Structural differences between proteins mainly arise from the variation in lengths and positioning of the loops.

Group IV: other α-helix proteins

There are seven families with very different functions in the 'other α helix' group (Figure 4). Skn-1 (1skn; Figure 4d) and MADS (see Figure 4g for the MADS box, 1mnm) are transcription regulatory regions in eukaryotic proteins, papillomavirus-1 E2 (2bop; Figure 4a) and EBNA1 (1b3t; Figure 4c) are viral transcription regulators and replication initiators, histones (1aoi; Figure 4b) and high-mobility group (HMG) proteins (1qrv; Figure 4f) are architectural proteins for DNA packaging, and Cre (1crx; Figure 4e) is a site-specific recombinase. Although the protein structures are very different, all use α helices (colored red in Figure 4) as the main method of DNA binding.

The Skn-1 and MADS proteins bind long probe helices in the DNA major groove in a manner similar to zipper-type proteins. Skn-1 (Figure 4d) is monomeric with a compact four-helix unit; the longest α helix at the carboxy-terminal end binds the major groove, and the rest of the domain contacts the DNA backbone [39]. In MADS (Figure 4g), an anti-parallel β sheet and an adjacent coiled-coil provide the dimerization interface. The α helices on the opposite face of the sheet diverge from the center of the binding site into adjacent major grooves, contacting base and backbone groups. The DNA is bent towards the protein [40].

Papillomavirus-1 E2 and EBNA1 are structurally similar dimeric proteins that can be divided into two regions (Figure 4a,c). In the core region, four β strands from each subunit combine in an eight-stranded β barrel. The flanking DNA-binding regions project single α helices into the DNA major



21. Leucine zipper (2dgc) 22. Helix-loop-helix (1am9)

Figure 3
Group III, zipper-type proteins. Colors, numbers and names are used as in Figure 1.

groove symmetrically. As is apparent from their structures (Figure 4a,c), the binding orientations of the helices are very different in the two families [41,42].

Histone and HMG are multimeric proteins that bind DNA independent of base sequence. Histone (Figure 4b) is an octameric protein whose structure can be approximated to a cylinder. Each subunit comprises a bundles of three or four helices that pack against each other; the long DNA segment wraps around the circular edge of the protein. Neighboring α helices make extensive contacts with DNA backbone groups to stabilize the distortion, but none is inserted in the groove and there are few interactions with bases [43]. The HMG subunit comprises three α helices that are arranged in an L shape (Figure 4f). The first and second helices bind base and backbone groups from the minor groove and cause severe distortion in the DNA structure through intercalation of amino-acid side chains [44].

Finally, Cre (Figure 4e) is a dimeric protein. Each subunit consists of two structural domains that fold into complex helical bundles. Jointly the domains form a clamp around the DNA, inserting α helices into both the major and minor grooves [45].

Group V: the β-sheet proteins

In contrast to the proteins described so far, groups V and VI comprise proteins that use β-strand structures for DNA recognition and binding. Group V, which only contains the TATA box-binding protein family, is characterized by the use of a wide β sheet to bind the DNA (Figure 5).

TATA box-binding proteins are an essential component of multiprotein transcription initiator complexes that assemble on promoters of genes transcribed by RNA polymerase II. Although they are single-chain molecules, their structures are generally considered to consist of two pseudo-identical domains. A ten-stranded antiparallel β sheet joining the domains covers the DNA minor groove; it creates two substantial kinks away from the main body of the protein by intercalating phenylalanine side chains from either end of the sheet [46,47]. The family is represented by proteins from the bacterium *Pyrococcus woesei*, yeast and humans. Both sequence and structural alignments of the various subunits yield very high SSAP scores (>90% and >90 respectively).

Group VI: the β-hairpin/ribbon proteins

The members of this group are different from the TATA box-binding proteins in that they use smaller two- or three-stranded β sheets or hairpin motifs to bind in either the DNA major or minor grooves (Figure 6). Six protein families of very diverse function are represented: the MetJ repressor (1cma; Figure 6a), Arc repressor (1bdt; Figure 6f) and T-domain families (1xbr; Figure 6d) constitute DNA-binding regions of transcriptional regulators; the integration host factor (1ihf; Figure 6c) and the hyperthermophile chromosomal proteins (1azp; Figure 6e) act as scaffolds to dictate the DNA structure for formation of high-order protein-DNA complexes; and the Tus protein (1ecr; Figure 6b) terminates DNA replication by helicases. Although the overall structures of the proteins are different, there are common themes in the use of the β strands.

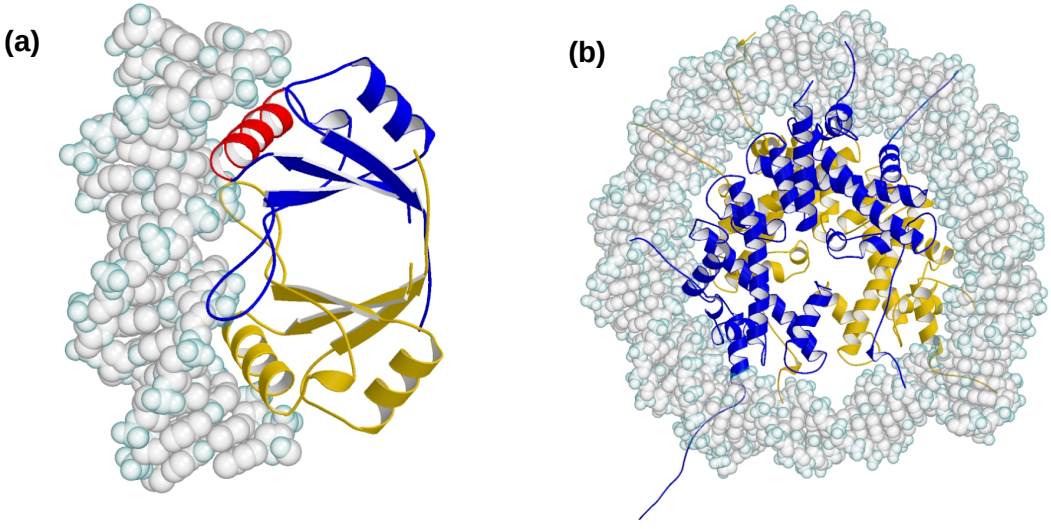
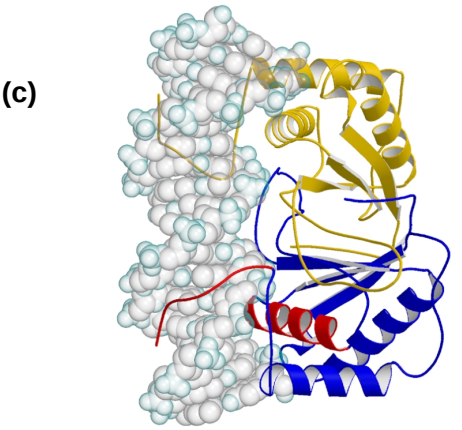
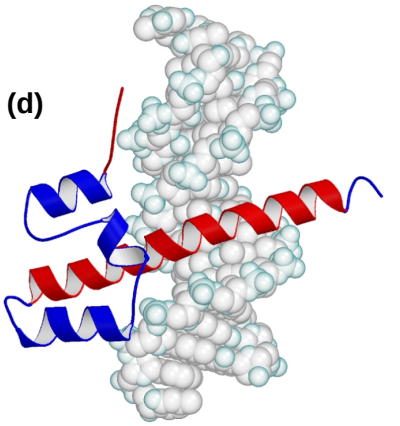


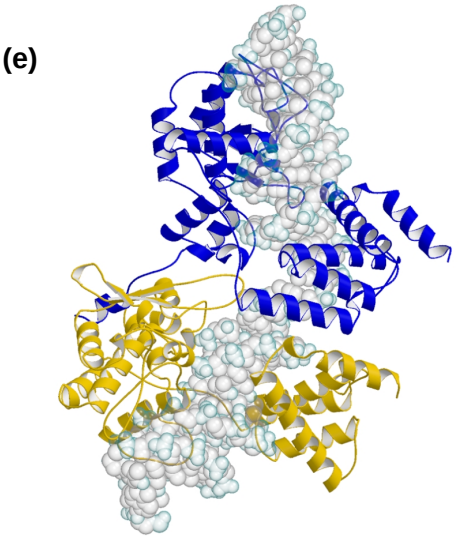
Figure 4 *(continues overleaf)*
Group IV, 'other α helix proteins'. Colors, numbers and names are used as in Figure 1.



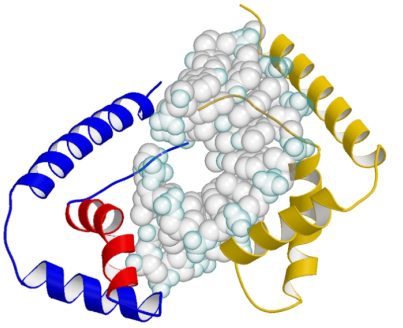
25. Ebna-1 nuclear protein (1b3t)



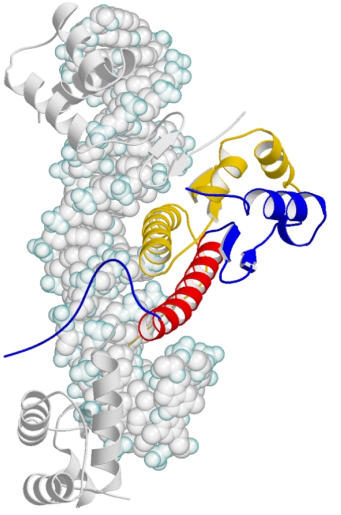
26. Skn-1 transcription factor (1skn)



27. Cre recombinase (1crx)



28. High mobility group (1qrv)



29. MADS box (1mnm)

Figure 4 *(continued)*
Group IV, 'other α helix proteins'.

The MetJ and Arc repressors are both dimers with very similar modes of binding (Figure 6a,f). Each protein subunit comprises a helical bundle and a single β strand; the strands from each subunit pack side by side, forming an antiparallel sheet that binds in the DNA major groove. The sheets lie flat in the groove; therefore protein side chains from just one face of the strand interact with base edges [48-50].

The Tus replication terminator and T-domain proteins use β-strand motifs to bind the DNA major groove (Figure 6b,d). In both, the strands are positioned almost perpendicular to the base edges, enabling contacts from amino acids that expose their side chains on either face of the sheet. The Tus replication terminator is a monomeric protein made of amino- and carboxy-terminal α-helical bundles that are connected by antiparallel β strands. The structure forms a large cleft in which the DNA is bound with the major groove facing the strands [50]. In contrast, the T-domain binds as a dimer. Each subunit consists of a β barrel: one end of the barrel points towards the DNA and presents two β strands, one of which extends into the major groove [51].

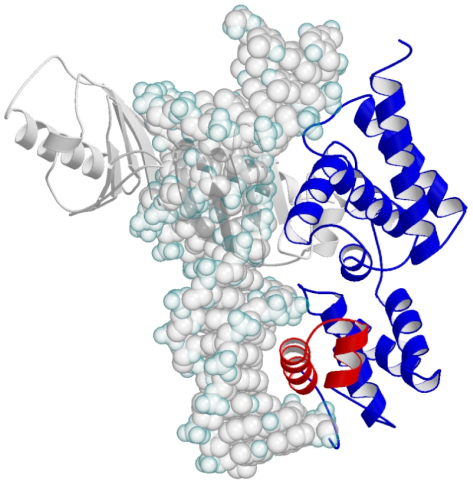
Both the integration host factor and chromosomal protein bind in the minor groove and distort the DNA by intercalating side chains from the β sheet motifs (Figure 6c,e). The integration host factor acts as a dimer; a β-hairpin arm from each subunit extends towards the opposing face of the DNA and inserts proline side chains between distinct base-steps [52]. The minor groove is widened in the region of binding and the DNA bends toward the main body of the protein. In contrast, the hyperthermophile chromosomal protein acts as a monomer and uses a three-stranded β sheet to bind against the minor groove. Two hydrophobic side chains from neighboring strands intercalate at a single base-step, causing the DNA to bend away from the protein [53].

Only the chromosomal protein and Arc repressor families contain more than one structure. Pairwise sequence identities and SSAP scores between subunits within families are high (>90% and >90 respectively).

Group VII: other

There are two non-enzymatic families in the current dataset that do not use a well-defined secondary structural motif for DNA binding (Figure 7). Both function as dimers and have multidomain subunits that mediate DNA-binding, dimerization and localization to the nucleus. Unlike the dimeric transcription factors encountered so far, these proteins envelop the nucleic acid and the complexes are symmetrical when viewed parallel to the DNA long axis. Interstrand and interdomain loops provide most of the base and backbone contacts.

The Rel homology region (Figure 7a) is a conserved amino-terminal domain of transcriptional regulators involved in



30. TATA box-binding family (1ytb)

Figure 5
Group V, β-sheet proteins. Colors, numbers and names are are used as in Figure 1.

cellular defense and differentiation. Each subunit comprises two β-sandwich domains, which are joined by up to ten interstrand loops that bind in the DNA major groove [54]. The STAT family (Figure 7b) contains transcription factors that mediate responses to cytokines and growth factors. Each protein subunit consists of four structural domains and the functional dimer resembles a pair of pliers with the DNA bound at the hinge. Surrounding loops and an α helix approach the DNA from both the major and minor grooves [55].

Group VIII: the enzymes

The enzyme group completes the classification of the dataset (Figure 8). Rather than having a common structural motif for binding DNA, proteins in the enzyme group are brought together on the basis of their functions; all alter DNA structure through the catalysis of a chemical process.

Unlike the proteins met with so far, the DNA-binding regions used by enzymes are generally hard to describe in terms of simple structural motifs, and these proteins use an extensive combination of α helices, β strands and loops to recognize and bind DNA (Figure 8). As described in the 'Outline of the families' section, below, many enzymes comprise three distinct domains: a DNA-recognition domain that 'reads' the DNA sequence; a catalytic domain with the enzymatic active site; and, where applicable, a dimerization domain, although clearly there are exceptions. The resulting structure often has a U-shaped cavity in which the DNA is bound [56] and often the DNA structure is severely deformed on binding.

For sequence-specific enzymes, the target sequences are typically 4-8 bp long, and binding is far more discriminating than that of the transcription regulators. For example, in proteins such as *Hha*I methyltransferases and endonucleases, a single change in the target sequence can lead to over a million-fold reduction in binding affinity. Proteins

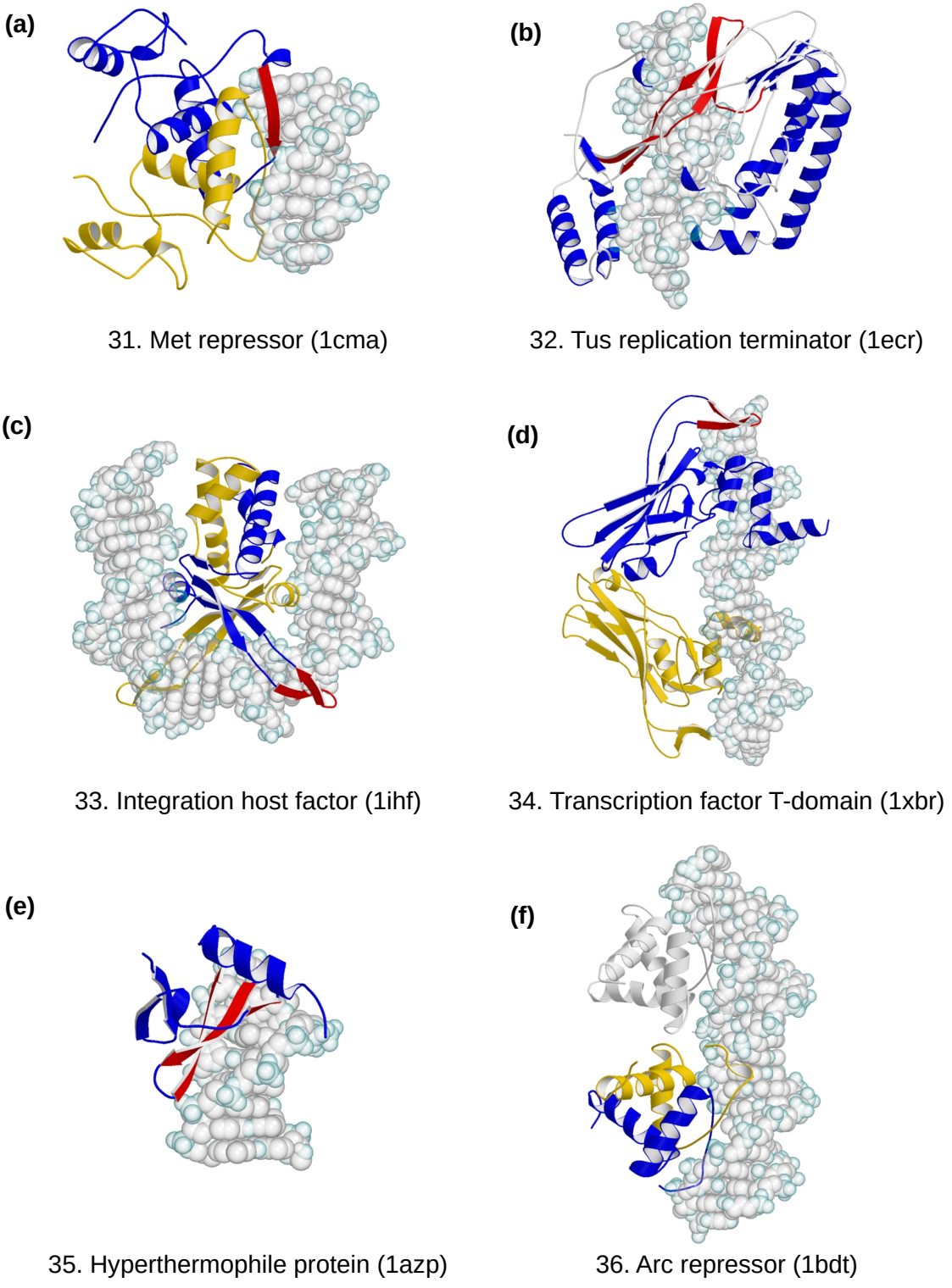


Figure 6
Group VI, the β-hairpin/ribbon proteins. Colors, numbers and names are used as in Figure 1.

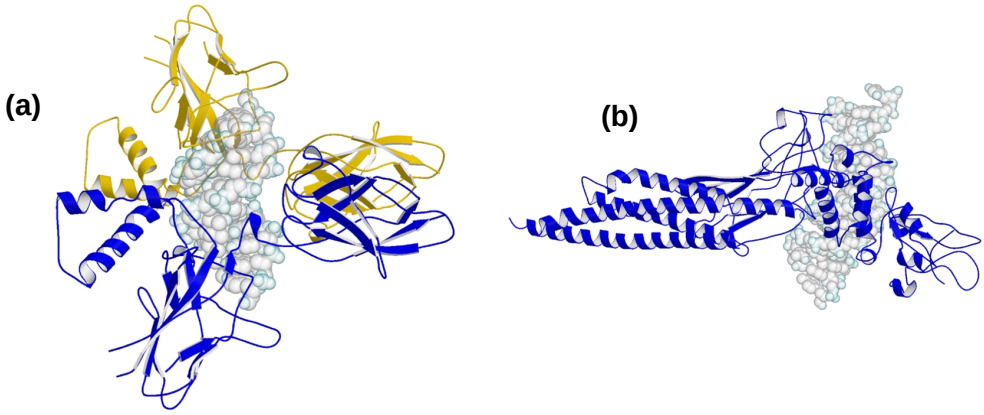


Figure 7
Group VII, 'other DNA-binding proteins'. Colors, numbers and names are used as in Figure 1.

are thought to derive their specificity from both read-out of the base sequence and the catalytic action on the DNA, as in endonucleases *Bam*HI (3bam; Figure 8e) and *Eco*RI (1qps; Figure 8d), or even primarily from the catalytic process, as in endonuclease *Eco*RV (1rva; Figure 8c) [57]. Other proteins, such as polymerases, must, however, provide sequence-independent interactions with their DNA substrate yet retain the specificity to distinguish correctly paired bases from mismatches. Seven endonucleases and four polymerases (see families 40-46 and 47-50, and Figures 8b-h and 8i-l, respectively), dominate this group of 16 families.

A protein-DNA complex website

A website that summarizes the groups and families of protein-DNA complexes can be found at [http://www.biochem.ucl.ac.uk/bsm/prot_dna/prot_dna.html]. The pages include a brief description of each family, similar to those given in the online information published with this review, as well as information on the aligned subunits of each protein, structural alignments, tables of pairwise sequence identities and SSAP scores. The proteins are linked to their respective PDB and NDB entries and a PRINTS [58] sequence motif analysis. Also available are links to PDBsum [59], our database of summaries and structural analyses of PDB data files. Each structure has information on its CATH [60], PROCHECK [61] and PROMOTIF [62] analyses and links to SCOP [63], WHATIF [64] check and FSSP [65] structural alignments.

The classification process will be automated in the near future so that a newly solved protein structure can submitted to the website and either grouped into an existing family or identified as novel. This would facilitate the possibility of being able to predict a DNA-binding motif and its binding site given a protein sequence, or pave the way to designing proteins to bind a given DNA sequence.

Conclusions

This data collection provides the basis for improving our understanding of protein-DNA complex formation. It highlights the diversity of protein-DNA complex geometries found in nature, but also underlines the importance of interactions between α helices and the major groove, which is the main method of binding in 28 of the 54 families. In particular, the HTH and zinc-coordinating motifs are used repeatedly, and provide compact frameworks that present the α helix on the surfaces of structurally diverse proteins, ready for interaction with the DNA. These structures show many variations, both in amino-acid sequence and detailed geometry, and have clearly evolved independently in accordance with the requirements of the contexts in which they are found. While achieving a close fit between the α helix and major groove, there is enough flexibility to allow both the protein and DNA to adopt distinct conformations, resulting in multispecific complementarity. Even for this interaction there does not appear to be a simple code relating amino-acid sequence to the DNA sequence it binds. Given the additional complexities of totally different frameworks, it is now clear that detailed rules for DNA base recognition will be family specific, but with underlying trends such as the arginine-guanine interactions.

This survey also highlights the differences between protein domains that 'just' bind DNA and those involved in catalysis. Although there are exceptions, the former typically approach the DNA from a single face and slot into the grooves to interact with the base edges. The latter commonly envelop the substrate using complex networks of secondary structures and loops, often causing significant distortions in the DNA - normally a requirement for the catalytic process. The ability to bend DNA is not only limited to the enzymes, however; although not as severe, DNA bending is clearly also a common feature of complexes formed by transcription factors. This and other

effects such as electrostatic, water- and cation-mediated interactions assist indirect recognition of the DNA sequence, although they are not well understood yet.

Of interest is how the current dataset will aid our interpretation of genome sequences. As summarized in Table 1, there are more structures of eukaryotic proteins than of prokaryotic, and very few are viral. It also demonstrates that, although the dataset is still limited, eukaryotic DNA-binding domains have greater structural diversity than others. This is unsurprising, given that these organisms have developed relatively sophisticated transcription and DNA-repair mechanisms, and therefore more eukaryotic proteins are likely to be found and to be structurally characterized. Although preliminary studies of the available genomes show that many proteins will probably fall into existing families - notably those with HTH, zipper-type and ββα zinc-finger motifs - there are exciting possibilities of discovering further modes of DNA-binding. Genome analysis will not only facilitate identification of such proteins, but will allow us to determine functionally important target sites on the DNA and, in combination with structural data, how higher-order oligomers are formed within the cell. Ultimately, this will expand our understanding of the regulation of protein expression and DNA packaging, rearrangement, repair and replication, which are indispensable to the viability of organisms.

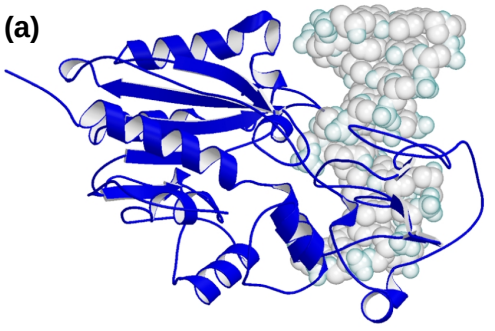
Outline of the families of DNA-binding proteins

A complete outline of the families of DNA-binding proteins and their functional, structural and binding properties follows. Box 1 shows the selection process by which the dataset was compiled. Table 1 provides a summary of the families and Table 2 lists the 240 structures of protein-DNA complexes in the database. Figures 1-8 show ribbon diagrams of the relevant structures.

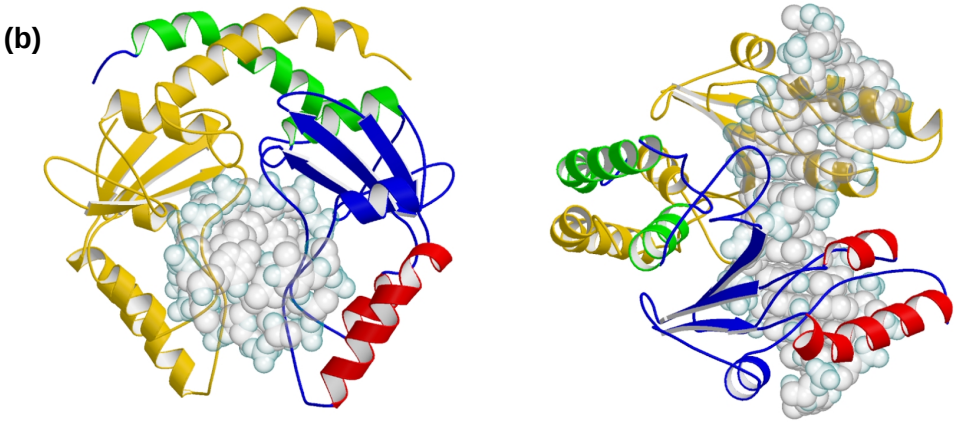
Group I: helix-turn-helix group

1. Cro and Repressor family

Function. The Cro and Repressor proteins (Figure 1a) are part of the lysogenic/lytic growth switch mechanism in bacteriophages and function as transcriptional regulators at a set of six related operons.



39. Methyltransferase family (6mht)



40. Endonuclease PvuII family (3pvi)

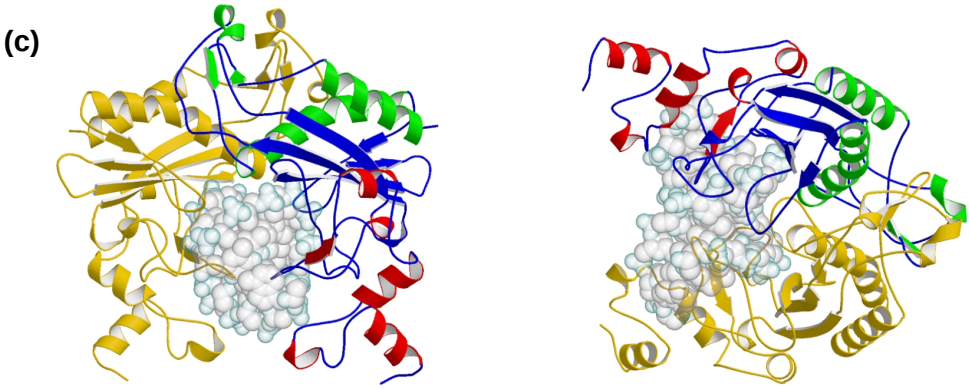
Figure 8 *(continues overleaf)*
Group VIII, the enzymes. Colors, numbers and names are used as in Figure 1.

Structure. Both protein types function as homodimers. Each Repressor subunit has two domains: an amino-terminal five-helix bundle whose second and third α helices comprise a HTH motif; and a carboxy-terminal domain that mediates dimerization (Figure 1a). Cro is a single-domain protein with a structure homologous to the amino-terminal region of Repressor. The fourth and fifth α helices mediate dimerization [66].

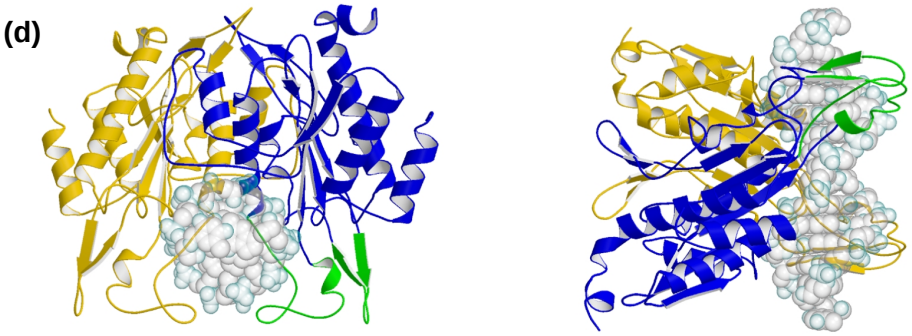
Binding. Cro and Repressor bind six related operons with varying affinities. Each operon is 14 bp long and pseudo-symmetrical; four bases at either end are conserved between sites and the variation in the sequence of the central 6 bp are thought to modulate the binding affinity of the protein. The recognition helix of the HTH motif contacts base edges in the DNA major groove.

2. Homeodomain family

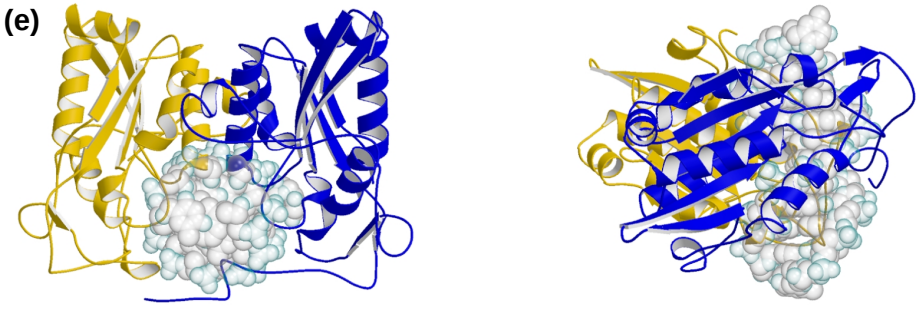
Function. These are transcription regulators for a wide range of genes; in particular many have a vital role in development and cell differentiation (for example, Mat α-2; 1apl). Some are expressed broadly whereas others are tissue-specific.



41. Endonuclease EcoRV family (1rva)



42. Endonuclease EcoRI family (1qps)



43. Endonuclease BamHI family (3bam)

Figure 8 *(continued)*
Group VIII, the enzymes.

Structure. The proteins are small (just over 100 amino-acid residues in length) and consist of four helices.

Binding. The protein binds DNA either as a monomer or a dimer, depending on the protein and many are capable of both. Typical HTH binding is displayed in Figure 1b, with the second helix of the motif inserted in the DNA major groove.

3. LacI repressor family

Function. Lac repressor regulates the lac operon, which

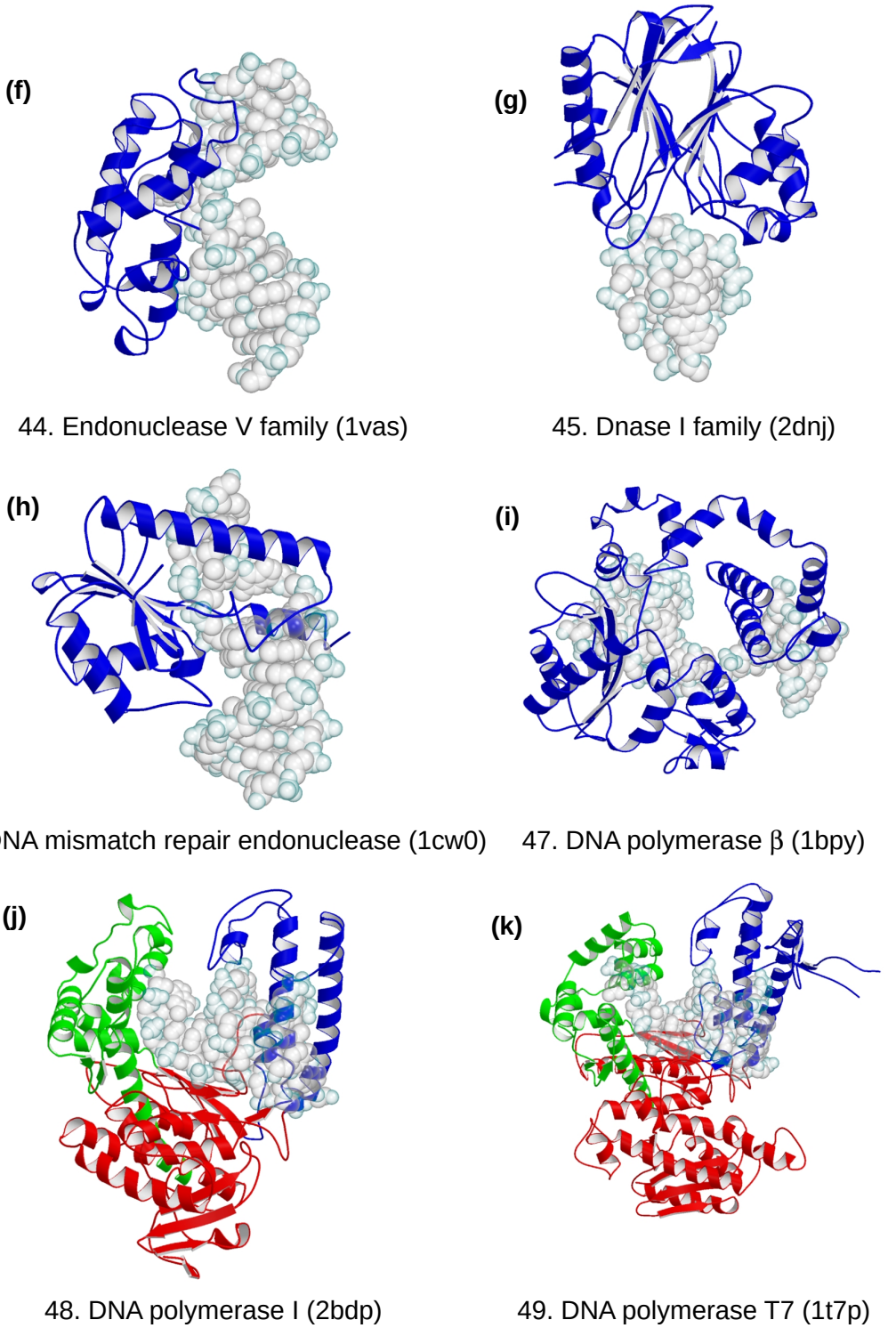
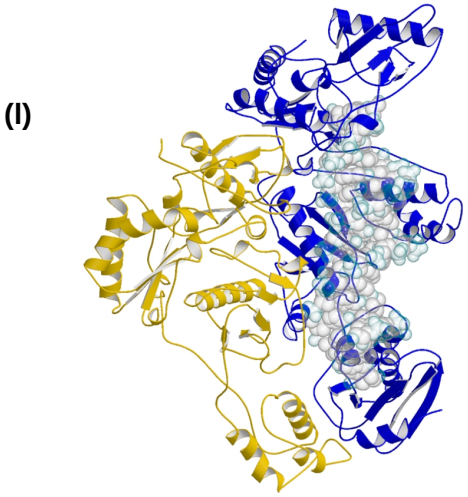


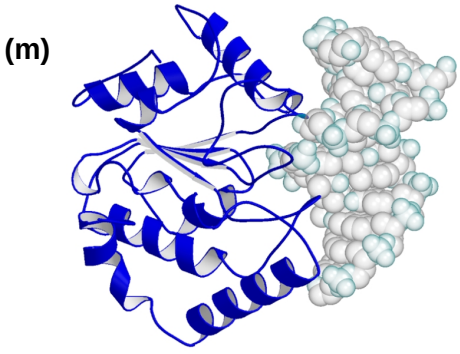
Figure 8 *(continued)*
Group VIII, the enzymes.

codes for proteins required to transport and degrade lactose. The purine repressor proteins of the LacI repressor family regulate *de novo* purine and pyrimidine synthesis by repression of genes encoding enzymes that participate in the synthesis pathway. Guanine and hypoxanthine act as co-repressors on binding to the protein. Other members of the LacI repressor family, not represented in the current dataset, display high structural and sequence similarity and control a wide range of biosynthetic pathways [21].

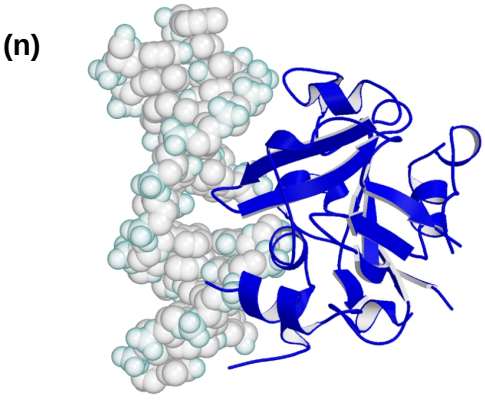
Structure. Purine repressors function as homodimers, as do most other family members (Figure 1c). The lactose, fructose



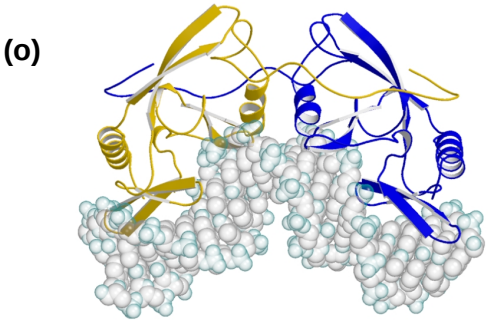
50. HIV reverse transcriptase (1hmi)



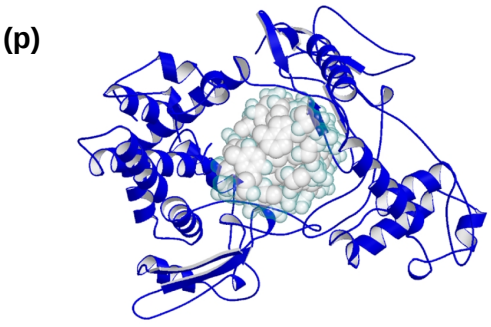
51. Uracil-DNA glycosylase (1ssp)



52. 3-Methyladenine DNA glycosylase (1bnk)



53. Homing endonuclease (1a73)



54. Topoisomerase I (1a31)

Figure 8 *(continued)*
Group VIII, the enzymes.

and raffinose repressors are exceptions, and appear to exist as tetramers [67]. Each subunit is a two-domain structure. The amino-terminal domain (approximately 60 residues) contains a three-helix bundle followed by a loop and an additional helix. The first two α helices form the HTH motif and the fourth is called the hinge helix. The larger carboxy-terminal domain (about 280 residues) is a mixture of α helices and β strands and binds the co-repressor.

Binding. Binding sites are typically 16-18 bp long and pseudo-palindromic. The recognition helix of the HTH motif binds in the major groove and phosphate backbone contacts are mediated by the remainder of the helical bundle. The hinge helix from each subunit is inserted in the same DNA minor groove at the center of the binding site and jointly introduce a kink by intercalation of leucine sidechains [68,69].

4. Endonuclease *Fok*I family

Function. Endonuclease *Fok*I is a bipartite restriction enzyme which recognizes a specific DNA sequence and non-specifically cleaves at a position a short distance away.

Structure. The protein acts as a monomer with two functional regions (Figure 1d). The amino-terminal DNA-recognition region (about 390 residues) may be divided into three further subregions. D1, a roughly 160 residue subregion made of an amino-terminal arm, ten α helices and a two-stranded β sheet. Helices 5, 6 and 8 form a pseudo-HTH motif. Helices 5 and 6 lie on the same helical axis, jointly forming the first α helix, and helix 8 acts as the recognition helix. A subregion (D2), of about 110 residues, contains six α helices and a three-stranded β sheet with the α helices packing in a triangular formation and the second and fifth α helices arranged in an HTH-like manner. The turn is replaced by an extensive loop region - D3 - an approximately 80-residue segment containing five α helices and a three-stranded β sheet. The carboxy-terminal catalytic domain (about 180 residues) is made of a five-stranded β sheet flanked by seven α helices. The active site is situated on the first three β strands in the region [70].

Binding. Binding is to a site containing the sequence 5′-GGATG-3′ and staggered cleavage occurs 9 and 13 bp away from the target sequence. All base contacts to the recognition sequence are made by subregions D1 and D2. The amino-terminal arm and second α helix from D1 bind in the major groove and a loop preceding this recognition helix is found in the minor groove. The recognition helix from the HTH motif in D2 contacts the major groove. The catalytic region is positioned adjacent to the DNA-recognition region.

5. γδ-resolvase family

Function. The γδ-resolvase is a site-specific recombinase which converts negatively supercoiled circular DNA containing two directly repeated copies of the recombination site into two interlinked rings.

Structure. The protein functions as a homodimer (Figure 1e). Each subunit is made of two domains. The amino-terminal domain (about 120 residues) contains the catalytic center and the dimerization interface. It consists of a five-stranded β sheet flanked by three α helices on one side and a single α helix on the other. The longest α helix packs with its counterpart in the other subunit to stabilize the dimer. The carboxy-terminal domain (approximately 40 residues) is a three-helix bundle with the second and third α helices forming a HTH motif. An extended arm region (about 20 residues), comprising the carboxy-terminal half of the dimerization helix and a loop, connect the two domains.

Binding. Each 114 bp recombination region consists of three resolvase-binding sites, I, II and III. Each site binds a resolvase dimer and is made of an inverted repeat of a 12 bp recognition sequence with varying base sequence and spacing between the half-sites. The structure found in 1gdt (Figure 1e) is thought to represent the conformation found prior to the recombination process. Two main DNA-binding regions are found in each subunit. The recognition helix of the carboxy-terminal HTH motif binds in the major groove at the outer ends of the binding site. The extended helix in the arm region is inserted in the minor groove near the center of the binding site in a similar manner to the recognition helices from leucine-zipper structures. The DNA is bent 60° away from the main body of the protein. The DNA is slightly kinked at the center of the site owing to partial intercalation of threonine residues from the arm region [71].

6. Hin recombinase family

Function. The Hin recombinase protein catalyzes site-specific recombination in the *Salmonella* chromosome.

Structure. The structure 1hcr is of the domain involved in DNA sequence recognition (Figure 1f). It is a three-helix bundle flanked by short peptide chains at either end (about 50 residues). The second two α helices form the HTH motif [18].

Binding. The full protein cooperatively binds as a homodimer at a 26 bp site. The recognition helix in the HTH motif is inserted in the major groove and surrounding helices make contacts with the phosphate backbone. The amino- and carboxy-terminal tails bind in adjacent minor grooves although their importance in sequence recognition is unknown.

7. RAP1 family

Function. The RAP1 protein performs two functions. The first is the periodic binding of DNA to regulate telomere length. Telomeres are nucleoprotein complexes found at the ends of eukaryotic chromosomes where the DNA consists of a repeated array of short, species-specific sequence motifs. The second function is that of transcription regulation; RAP1 functions as an activator or repressor for a large number of genes.

Structure. RAP1 is a monomeric protein with two homologous domains and a carboxy-terminal tail (Figure 1g). Domain 1 (about 80 residues) contains a three-helix bundle and an amino-terminal tail, whereas domain 2 (about 80 residues) contains an additional fourth α helix. In each, the second and third α helices form the HTH motif. The two domains are connected by a 30 residue linker region and are positioned 8 bp apart. The carboxy-terminal tail is a 20 residue segment which emerges from domain 2 and folds back towards domain 1 [72].

Binding. The binding site is 16 bp long and shows a tandem repeat at an 8 base interval. The two domains bind in a similar fashion at opposite ends of the binding site; the recognition helices of the HTH motif are inserted in the major groove and the remaining α helices contact the neighboring DNA backbone. The amino-terminal and the linker regions interact with the minor groove and the carboxy-terminal tail interacts with the major groove as it folds back. The flexibility of the linker allows for slight variations in spacing between tandem repeats.

8. Prd paired domain family

Function. The Prd paired domain is a functional domain found in a set of transcription regulatory proteins which are important in cell development.

Structure. The protein acts as a monomer with two structural domains (Figure 1h). The amino-terminal domain (about 70 residues) contains a short antiparallel β sheet and a β turn followed by a three-helix bundle and extended carboxy-terminal tail. The second and third α helices in the bundle form an HTH motif. The carboxy-terminal domain (approximately 50 residues) also contains a three-helix bundle which has an HTH motif [73].

Binding. Prd proteins bind to 13-20 bp sites which share a common core sequence. The recognition helix in the HTH and the β turn of the amino-terminal domain make base contacts in the major and minor grooves respectively. The rest of the domain interacts with the DNA backbone. The carboxy-terminal domain does not contact the DNA, but domain structure and biochemical evidence suggest it does bind DNA in certain family members (for example Pax proteins).

9. Tc3 transposase family

Function. The structure contained in 1tc3 (Figure 1i) is of the DNA-recognition domain found in the amino terminus of Tc3 transposase. The function of the enzymes is to move specific segments of DNA from one position of the genome to another.

Structure. The domain (about 50 residues) contains a three-helix bundle and an amino-terminal tail (Figure 1i). The last two α helices form the HTH motif [74].

Binding. Binding is to a 20 bp site. The recognition helix of the HTH motif is bound in the major groove and other α helices make DNA backbone contacts. The amino-terminal tail binds in an adjacent minor groove although the interactions are not thought to be specific.

10. Trp repressor family

Function. The Trp repressor is involved in the regulation of tryptophan synthesis by binding three different operator sites. L-tryptophan acts as co-repressor.

Structure. Each subunit (about 100 residues) forms a six-helix bundle (Figure 1j). Helices 4 and 5 correspond to the HTH motif whereas the remaining four α helices provide the dimerization interface. Tryptophan also binds the helical bundle [23].

Binding. Binding is to three related 16 bp operator sites which the protein binds in the presence of tryptophan. The HTH motifs are reoriented on binding of the co-repressor to enable DNA-binding. The recognition helix is positioned in the major groove and most base contacts are made through a network of intermediate water molecules. Operator sites are symmetrical and also show approximate symmetry within the half-site, which leads to two alternative modes of binding. In the first, the dimer subunits bind each half-site symmetrically about the central base-pairs. This is similar to what is observed for the other prokaryotic HTH proteins. In the second, two dimers co-operatively bind to a single operator site in tandem. Dimers are staggered by 8 bp and rotated through 270° about the DNA axis and the crystal structure 1trr (Figure 1j) displays a superhelix of dimers binding successive binding sites.

11. Diphtheria tox repressor family

Function. The virulent phenotype of the pathogenic bacterium *Corynebacterium diphtheriae* is conferred by diphtheria toxin, whose expression is an adaptive response to low concentrations of iron. The expression of the toxin gene (*tox*) is regulated by the repressor diphtheria tox, which is activated by transition metal ions.

Structure. Diphtheria tox is a 225-residue protein that binds as a dimer to DNA. Each monomer consists of six helices and a short two-stranded β sheet, with helices 2 and 3 constituting the HTH motif (Figure 1k).

Binding. The DNA interacts with two dimers bound to opposite sides of the *tox* operator, with each dimer interacting with two major groove regions. Together, the two HTH motifs (one in each dimer) bind a 24 bp sequence.

12. Transcription factor TFIIB

Function. The transcription factor TFIIB is an essential part of the multiprotein transcription initiator complex that assembles on RNA polymerase II promoters. TFIIB

binds a 7 bp region upstream of the TATA box called the B recognition box.

Structure. TFIIB is composed almost entirely of α helices and is approximately 200 residues long (Figure 1l).

Binding. TFIIB binds DNA in two places as a result of the nucleic acid distortion caused by the interaction of the TATA box-binding protein. The main interactions are due to a carboxy-terminal HTH motif binding DNA in the major groove at the upstream site. The protein also binds DNA in the minor groove at a downstream site using the amino terminus of a helix to contact the DNA backbone.

'Winged' HTH proteins

13. Interferon regulatory factor family

Function. The family of interferon regulatory factor (IRF) transcription factors is important in the regulation of interferons in response to infection by virus and in the regulation of interferon-inducible genes.

Structure. The IRF family is characterized by a unique 'tryptophan cluster' DNA-binding region of five tryptophan residues. The protein binds as a monomer with a HTH motif binding DNA through three of the five conserved tryptophans. The IRF DNA-binding region has an α/β architecture consisting of a cluster of three α helices flanked on one side by a mixed four-stranded β sheet (Figure 1m).

Binding. Helices 2 and 3 comprise the HTH motif, with helix 3 lying in the DNA major groove. Contacts to bases within the major groove are localized to a GAAA core sequence within a 13 bp DNA element in the interferon promoter.

14. Catabolite gene activator (CAP) family

Function. CAP is a cAMP-dependent transcription regulator. A rise in cAMP concentration leads to increased affinity of CAP for catabolite-sensitive operons.

Structure. The protein functions as a homodimer, and each subunit comprises a two-domain structure (Figure 1n). The carboxy-terminal domain (about 60 residues) mainly consists of a three-helix bundle with the second two α helices forming the HTH motif. The domain contains a small β sheet that also contributes to DNA binding. The larger amino-terminal domain (approximately 130 residues) has an extensive β sheet that mediates cAMP binding, and a long α helix that forms the dimer interface [75].

Binding. The consensus binding sequence is a symmetric 22 bp site. Binding by the recognition helix of the HTH motif in the major groove induces a sharp, highly localized bend in the DNA and additional contacts with the phosphate backbone are made by the β strands from the same domain.

15. Transcription factor family
Heat-shock and E2F/DP transcription factors

Function. The protein 3hts (Figure 1o) recognizes the promoters of the heat-shock protein genes through upstream DNA sequences (heat-shock elements, HSEs). An HSE consists of alternating, inverted repeats of the sequence nGAAn, where n can be any nucleotide. The E2F and DP protein families form heterodimeric transcription factors that have a central role in the expression of cell-cycle-regulated genes and recognize a c/gGCGCg/c sequence.

Structure. The DNA-binding domains of these proteins have a 'winged' HTH fold - that is, a three-helix bundle capped by an antiparallel β sheet. Helices 2 and 3 constitute the HTH motif.

Binding. The third helix of the HTH is docked into the major groove. The DNA-binding domain makes additional contacts to the DNA through the amino terminus of the first helix and the turn of the HTH motif. The only other HTH fold that contacts the DNA with the residues of the turn is the Ets family.

16. Ets domain family

Function. The Ets family of transcription factors, of which there are now about 35 members, regulate gene expression during growth and development. They share a conserved domain of around 85 amino acids which binds as a monomer to the DNA sequence 5′-C/AGGAA/T-3′.

Structure. The 'winged' HTH motif interacts with a 10 bp region of duplex DNA that takes up a uniform curve of 8° (Figure 1p).

Binding. The domain contacts the DNA by a loop-helix-loop architecture, the turn of the HTH motif and the loop at the end of helix 1 before the β sheet contacting the DNA backbone.

Group II: zinc-coordinating proteins

17. ββα zinc-finger family

Function. The ββα zinc-finger proteins constitute the largest individual family in this group. The DNA-binding motif is found in many transcription regulators and more than a thousand distinct motifs have been identified through sequence analysis [26].

Structure. The structure of the finger is characterized by a short two-stranded antiparallel β sheet followed by an α helix (Figure 2a) and a single zinc ion bound by two pairs of conserved histidine and cysteine residues situated in the α helix and second β strand. Proteins generally contain multiple copies of fingers in a single peptide chain which wrap round the DNA along the major groove in a spiral manner.

Binding. The recognition pattern of the probe α helix has been well characterized; each finger binds adjacent 3 bp subsites on the DNA using amino acids at positions -1, 2, 3 and 6 relative to the start of the α helix, -1 being the residue position preceding the helix [2,24,76]. Although exceptions to this rule have been observed in specific examples [29,77], experiments have shown that by altering the amino-acid types at the key positions, different subsite sequences are recognized, suggesting that these residue positions are usually sufficient for specific binding [30,78]. By varying the number of fingers used in a protein chain, this relatively simple motif allows recognition of a wide range of binding sites with different degrees of specificity. For example, a protein with five fingers is expected to bind a site very selectively, whereas a protein with only a single finger would bind a wide range of sites containing the required 3 bp sequence. However, the structure of the human glioblastoma protein suggests that binding is not always straightforward; of the five fingers in the structure, one does not contact the DNA at all and only two appear to make specific contacts with bases [31]. As described earlier, the protein subunits in this study have been split into distinct domains, each containing a single zinc-finger motif. The pairwise sequence identities of the aligned domains are all high, ranging from 73% (for example, human zinc-finger protein, 1udbA1, and *Drosophila* tramtrack protein, 2drpA1) to 100% (for example, mouse Zif268 protein, 1aayA1, and artificial protein, 1mey). All domains are structurally very similar, returning SSAP scores of over 90.

18. Hormone receptor family

Function. Members of the hormone receptor family translocate from the cytoplasm to the nucleus and regulate transcription at DNA sequences called hormone response elements on binding of steroid and other hormones [2,32].

Structure. Hormone receptors function as homo- or hetero- dimers and each monomer typically consists of a ligand-binding, a DNA-binding and a transcription regulatory domain (Figure 2b). The zinc-coordinating motif is found in the DNA-binding domain and is characterized by two antiparallel α helices capped by loops at their amino-terminal ends. Each helix-loop pair coordinates a single zinc ion using four conserved cysteines. The two α helices lie approximately at right angles to each other; the first is inserted in the DNA major groove to provide interactions with bases whereas the loops and the second α helix contact the DNA backbone. The DNA-binding domain alone is sufficient for dimerization, the interface being formed by the loops leading into the second α helix.

Binding. All receptor subunits bind to one of two half-site sequences, 5′-AGAACA-3′ or 5′-AGGTCA-3′. A hormone-response element contains two half-sites and the identity of the response element is determined by the sequences that are present, the relative orientation between them (either symmetric or palindromic) and the spacing between them (between 3 and 6 bp). Thus recognition of the target sequence by the whole hormone receptor depends on read-out of half-site sequences by each subunit and the structure of the homo- or heterodimeric protein [33]. The sequences of all subunits in the current dataset are very similar (sequence identities > 90%) except for the DNA-binding domain of the thyroid hormone receptor (for example, 1bsx), which has two extra helices in the carboxy-terminal tail. The structures are all very similar with pairwise SSAP scores of over 90.

19. Loop-sheet-helix family

Function. The loop-sheet-helix zinc-binding motif is represented solely by the DNA-binding region of p53, a transcriptional activator implicated in tumor suppression [2,34].

Structure. As the name indicates, the DNA-binding domain consists of a loop leading out of the main body of the protein, followed by a small β sheet, an α helix and then another loop that leads back into the protein (Figure 2c). The zinc ion is coordinated by three cysteines and a histidine in the two loop regions.

Binding. Base contacts are supplied by the α helix in the DNA major groove and by the loops in the minor groove, although the latter are not thought to confer much specificity. The protein functions as a tetramer, with each subunit contacting a separate 5 bp recognition sequence positioned one after another. All intersubunit interactions are made by regions outside the DNA-binding motif.

20. Gal4-type family

Function. The final zinc-coordinating family contains only the Gal4 protein [24,35]**.** It is a transcriptional regulator of galactose-induced genes and its zinc-coordinating motif has so far only been identified in proteins from *Saccharomyces cerevisiae*.

Structure. The motif consists of a pair of α helices that coordinate two zinc ions through six cysteine residues, where two of the cysteines are shared by both metal atoms (Figure 2d).

Binding. The first α helix is presented in the DNA major groove for binding with bases, and backbone interactions are made by the second α helix. Gal4 functions as a homodimer and the dimerization interface is located outside the zinc-coordinating motif.

Group III: zipper-type proteins

21. Leucine zipper family

Function. The leucine zipper family consists of the yeast GCN4 proteins that bind promoter regions of genes encoding enzymes involved in amino-acid biosynthesis, and the

Fos-Jun heterodimer, which activates the expression of many immune-response genes.

Structure. The structure of the zipper-type proteins may be split into two parts: the dimerization and DNA-binding regions. As shown in (Figure 3a), each subunit in the leucine zipper protein consists of a single α helix about 60 amino acids long. Dimerization is mediated through the formation of a coiled coil by a section of 30 amino acids at the carboxy-terminal end of each helix. The segment, known as the zipper region, consists of leucine or a similar hydrophobic amino acid every eight residue positions, roughly every two turns of the α helix. Corresponding side chains from each subunit mediate hydrophobic contacts at the interface through side-by-side packing. The DNA-binding region, also known as the basic region, is found in the amino terminus, and for the leucine zipper proteins, the binding segment is a direct extension of the dimerization region.

Binding. The α helices of the two subunits diverge from the coiled coil and enter the DNA major groove in opposing directions, each binding to half of the target sequence [2,36,79].

22. Helix-loop-helix family

Function. The helix-loop-helix proteins are transcription factors that control the expression of a wide range of genes involved in differentiation and development.

Structure. As the name suggests, helix-loop-helix proteins are a modification of the continuous α helices of the leucine zipper proteins in which the DNA-binding and dimerization regions are separated by a loop, resulting in a four-helix bundle (Figure 3b).

Binding. Like the leucine zippers, the dimerization helices interact with each other in a coiled-coil arrangement and the DNA-binding helices are inserted into the DNA major groove. By separating the two segments, more flexibility is allowed in positioning the probe helices on binding nucleic [2,37,38].

The helix-loop-helix family is represented by the mouse and human forms of Max, Srebp-1, mouse MyoD and human USF proteins. Sequence identities range from 66% (Max protein, 1an2A, and USF protein, 1an4A) to 97% (mouse Max protein, 1an2A, and human Max protein, 1hloA) and with the exception of the MyoD (1mdyA) and USF (1an4A) protein pair (pairwise SSAP score 70), SSAP scores are above 80. Structural differences between proteins mainly arise from the variation in lengths and positioning of the loops.

Group IV: other α-helix proteins

23. Papillomavirus-1 E2 family

Function. This family has a single member, the papillomavirus-1 E2 protein, which uses a probe helix as part of the DNA-recognition domain. The protein is a viral transcription regulator that acts at all viral promoters and also functions as a viral replication initiator.

Structure. The DNA-binding region of the E2 protein (Figure 4a) is about 85 residues long and consists of four β strands and two interstrand α helices. Two subunits combine to form an eight-strand β-barrel, which provides the interface for the resulting homodimer.

Binding. The larger α helix from each subunit is symmetrically inserted in the DNA major groove making base and backbone contacts. Additional interactions to the backbone are provided by interstrand loops [41].

24. Histone family

Function. DNA in chromatin is organized in arrays of nucleosomes. The nucleosome, in its role as the principal packaging element of DNA within the nucleus, is the primary determinant of DNA accessibility.

Structure. Two copies of each of four histone proteins are assembled into an octamer that has 145-147 bp of DNA wrapped in a superhelix around it to form a nucleosome core.

Binding. The protein octamer is divided into four 'histone-fold' dimers, each dimer being defined by H3-H4 and H2A-H2B histone pairs. The central histone-fold domains of all four core histone proteins share a highly similar structural motif constructed from three α helices connected by two loops. The two H3-H4 pairs interact through a four-helix bundle formed only from the two H3 histone folds to define the H3-H4 tetramer. Each H2A-H2B pair interacts with this tetramer through a second, homologous four-helix bundle between H2B and H4 histone folds. The histone-fold regions of each tetramer bind to the center of the DNA, which is wrapped into a superhelix. Further α helices and coil elements extend from the histone-fold regions and are also an integral part of the core protein within the confines of the DNA superhelix.

25. EBNA1 protein (Epstein-Barr nuclear antigen 1)

Function. EBNA1 binds to four recognition sites in the origin of latent DNA replication of Epstein-Barr virus and activates latent-phase replication of the viral genomes.

Structure. EBNA1 comprises two domains (Figure 4c), a flanking and a core domain (which is structurally homologous to the complete DNA-binding domain of the bovine papilloma virus E2 protein) and binds DNA as a dimer.

Binding. The flanking domain, which includes a helix that projects into the major groove and an extended chain that travels along the minor groove, makes all of the sequence-determining contacts with the DNA. The core domain makes no direct contacts with the DNA bases.

26. Skn-1

Function. Skn-1 is a developmental transcription factor that specifies mesoderm in *Caenorhabditis elegans*.

Structure. Skn-1 consists of a compact four-helix unit with one helix more than twice as long as any of the others (Figure 4d).

Binding. It binds as a monomer and binds DNA at two contact points. At the carboxy terminus, the longest helix extends from the domain to occupy the major groove of DNA in a manner similar to zipper proteins. Skn-1, however, lacks the leucine zipper found in all zipper. Additional contacts with the DNA are made by a short basic segment at the amino terminus of the domain, reminiscent of the 'homeodomain arm'.

27. Cre recombinase family

Function.: Cre recombinase catalyzes a site-specific recombination reaction between two 34-bp *loxA* and *loxP* sites in bacteriophage λ.

Structure. Cre is a 320-residue protein and folds into two distinct domains that are separated by a short linker. The amino-terminal domain contains five helices and the large carboxy-terminal domain is primary α helical with a small β sheet packing against a nine-helix domain (Figure 4e).

Binding. The protein binds DNA as a dimer, each monomer binding the outermost 15 bp of one lox half-site. The amino- and carboxy-terminal domains form a clamp around the half-sites making extensive contacts with both major and minor grooves. Helices 2 and 4 of the amino-terminal domain cross each other, both contacting the major groove of the *lox* half-site. The interface of DNA with the carboxy-terminal domain is complex, involving the entire face of the domain, with both helices and connecting loops interacting with the major and minor grooves and the DNA backbone.

28. High-mobility group family

Function. The high-mobility group (HMG) chromosomal proteins, which are common to all eukaryotes, bind DNA in a non-sequence-specific fashion to promote chromatin function and gene regulation. They interact directly with nucleosomes and are believed to be modulators of chromatin structure. They are also important in activating a number of regulators of gene expression, including p53, Hox transcription factors and steroid hormone receptors, by increasing their affinity for DNA.

Structure. Chromosomal HMG proteins have a global fold of three helices stabilized in an 'L-shaped' configuration by two hydrophobic cores (Figure 4f).

Binding. The HMG domain binds to an AT-rich DNA sequence using a large surface on the concave face of the protein, to bind the minor groove of the DNA. This bends the DNA helix axis away from the site of contact. The first and second helices contact the DNA, their amino termini fitting into the minor groove, whereas helix 3 is primarily exposed to solvent. Partial intercalation of aliphatic and aromatic residues in helix 2 occurs in the minor groove.

29. MADS-box family

Function. The MADS-box motif is found in various DNA-binding proteins, commonly transcription factors, and specifies DNA binding, dimerization and interaction with accessory factors.

Structure. MADS proteins bind DNA as dimers as part of a larger cooperative DNA-binding complex containing other DNA-binding proteins. The MADS domain is a 56-residue motif consisting of a pair of antiparallel coiled-coil α helices packed against an antiparallel two-stranded β sheet. This β sheet of the motif is also involved in inter-protein interactions with other accessory proteins.

Binding. MADS dimerization occurs along the extensive flat side of the monomer involving the helices and β sheet. The MADS protein shown here, MCM-1 (Figure 4g), interacts with DNA predominantly with its long α helices located nearly parallel to the minor groove at the center of the binding site. These α helices extend into the major groove on either side of the dyad; direct contacts made within the major groove and along the phosphate backbone cause the DNA to bend around the MADS box. The amino-terminal strand of the MADS region (before the first helix of the MADS motif) often passes over and interacts with the DNA backbone.

Group V: β-sheet proteins

30. TATA box-binding family

This group, which only contains the TATA box-binding protein family, is characterized by the use a large β-sheet structures to bind the DNA (Figure 5).

Function. TATA box-binding proteins are an essential component of the multiprotein transcription initiator complex that assembles on promoters bound by RNA polymerase II.

Structure. Although they are single-chain molecules, their structures are generally considered to consist of two pseudo-identical domains. A ten-stranded antiparallel β sheet joins the domains.

Binding. The β sheet covers the DNA minor groove and creates two substantial kinks away from the main body of the protein, by intercalating phenylalanine side chains from either end of the sheet [46,47].

The family is represented by *Pyrococcus woesei, Saccharomyces cerevisiae* and human forms of the protein.

Unsurprisingly, both sequence and structural alignments of the various subunits yield very high scores (> 90% and ≥ 90 respectively).

Group VI: β-hairpin/ribbon proteins

31. MetJ repressor family

Function. Transcriptional regulator of the expression of methionine biosynthetic enzymes in *E. coli*.

Structure. The MetJ repressor binds DNA as a dimer (Figure 6a), each subunit comprising a helical bundle and a single β strand; the strands from each subunit form the antiparallel sheet for DNA-binding (colored red).

Binding.: The two β strands fit into the major groove and do not alter the DNA structure significantly on binding. They lie flat against the base of the groove and interactions are only made from one face of the sheet. Supporting backbone contacts are made by the surrounding helices and the amino-terminal loop regions [48].

32. Tus replication terminator family

Function. Tus protein terminates replication of DNA in *E. coli*.

Structure. The protein consists of two α-helical bundles at the amino and carboxy termini, connected by a large β-sheet region and binds DNA as a monomer.

Binding. The DNA-binding region of the Tus family is made of four antiparallel β strands (colored red in Figure 6b) which links the amino- and carboxy-terminal domains and produces a large central cleft in the protein. The DNA is bound in this cleft, with the interdomain β strands contacting bases in the major groove. DNA backbone contacts are provided by the whole protein. The β strands are positioned almost perpendicular to the base edges in the groove, enabling contacts from amino acids that expose their side chains on either face of the sheet [50].

33. Integration host factor family

Function. Integration host factor (IHF) is a small heterodimeric protein that specifically binds to DNA and functions as an architectural factor in many cellular processes in prokaryotes.

Structure. The protein is a heterodimer of two related subunits each made of three helices and a two-stranded β sheet.

Binding. In contrast to the two families above, the integration host factor forces an enormous distortion in the DNA by inserting a β hairpin from each subunit in the minor groove (red in Figure 6c). As seen in the TATA box-binding family, the protein produces kinks by intercalating side chains between base steps at the edges of the binding sites. The intercalating prolines are found at the tips of the β hairpins that extend from the protein towards the other side of the DNA. The nucleic acid is bent towards the main body of the protein and the deformation is stabilized by contacts with the phosphate groups [52,80].

34. T-domain family

Function. The T domain (Figure 6d) is an approximately 180-residue homodimeric domain found in transcriptional regulators for genes essential in tissue specification, morphogenesis and organogenesis.

Structure. Each subunit consists of a seven-strand antiparallel β barrel; one opening of this barrel forms a dimer interface with the equivalent segment of the other subunit while the other end points towards the DNA.

Binding. Two β strands protrude from the barrel, one of which extends into the DNA major groove. The probe helix is situated in a three-helix bundle in the carboxy-terminal tail. In contrast to many protein families, the α helix binds base and backbone groups from the DNA minor groove [51].

35. Hyperthermophile chromosomal proteins

Function. These proteins are found in hyperthermophilic archaebacteria and have high thermal, acid and chemical stability. They bind DNA without marked sequence preference and increase the T_m of DNA by about 40°C.

Structure. The proteins consist of an incomplete five-stranded β-barrel capped by an α helix abutting three β strands (Figure 6e).

Binding. The proteins bind the minor groove with the three-stranded β sheet causing the DNA to kink severely. The kink results from the intercalation of specific hydrophobic side chains into the DNA structure, but without causing any significant distortion of the protein structure relative to the uncomplexed protein in solution.

36. Arc repressor

Function. Transcription of the *ant* gene during lytic growth of bacteriophage P22 is regulated by the cooperative binding of two Arc repressor dimers to a 21-bp operator site.

Structure. Arc is a small (about 100 residues), homodimeric repressor of the ribbon-helix-helix family of transcription factors. Each monomer consists of a pair of helices connected by an antiparallel β sheet (Figure 6f).

Binding. Each Arc dimer uses the β sheet to recognize bases in the major groove and the amino termini of the second helix in each pair contact the DNA backbone.

Group VII: other

37. Rel homology region family

Function. The Rel homology region is found in the amino terminus of proteins that act at the κB DNA recognition site, and mediates DNA binding, dimerization and nuclear localization (Figure 7a). Proteins that contain the region act as transcription regulators for genes commonly involved in cellular defense and differentiation. The carboxy-terminal domains located outside the region are variable between proteins.

Structure. The Rel homology region binds symmetrically as a homo- or hetero- dimer. Each subunit (of about 300 residues) has two distinct domains, both consisting of a β sandwich.

Binding. Interactions in the DNA major groove are made along the whole length of the 10-bp site using a total of ten interstrand loops [54,81].

38. STAT protein family

Function. STATs are a family of eukaryotic transcription factors that mediate the response to a large number of cytokines and growth factors. Upon activation by cell-surface receptors or their associated kinases, Stat proteins dimerize, translocate to the nucleus and bind to specific promoter sequences.

Structure. STAT proteins are between 750 and 850 residues long and bind as dimers to DNA target sites with a 9 bp consensus sequence, TTCCGGGAA. Each monomer is composed of four domains: an amino-terminal four-helix bundle, an eight-stranded β barrel (residues 321-465), a helix-loop-helix 'connector' domain (residues 466-585) and an SH2 domain.

Binding. The STAT homodimer grips the DNA like a pair of pliers (Figure 7b). The monomers are held together by the carboxy-terminal SH2 domains, and the large four-helix bundle domains form the 'handles' of the pliers. The DNA is almost entirely enclosed by the protein dimer, and contacts the loops from the β barrel and the connector domains.

Group VIII: enzymes

39. Methyltransferase family

Function. The methyltransferase enzyme is represented by a single homologous family [82,83]. The protein catalyzes the transfer of a methyl group from *S*-adenosyl-L-methionine to the C5 position of cytosine. In prokaryotes the reaction is most commonly found in the protection of the DNA from restriction enzymes. In eukaryotes, however, DNA methylation is implicated in a wider range of cellular processes including transcriptional regulation, DNA repair, developmental regulation and chromatin organization. The current dataset only includes the prokaryotic *Hha*I methyltransferase (for example, 4mht).

Structure.: The protein functions as a monomer (about 320 residues) containing two domains that are separated by a large DNA-binding cleft (Figure 8a). The catalytic domain (about 220 residues) consists of a seven-stranded β sheet flanked by a total of five α helices on either side. This domain contains the cofactor-binding site and the active sites. The DNA-recognition domain (about 100 residues) comprises five antiparallel strands that form a twisted β sheet.

Binding. The protein preferentially binds the sequence 5′-GCGC-3′ with the first cytosine base methylated in the enzyme reaction. The DNA is bound in the protein cleft so that the major groove faces the recognition domain and the minor groove faces the catalytic domain. The 4 bp in the target sequence are contacted from the major groove using two glycine-rich interstrand loops, and the substrate cytosine is flipped out of the DNA helix into the catalytic domain. The DNA structure is underwound and the base-pairing is rearranged over 3 bp either side of the substrate base. The three structures in the family all have identical sequences and return high pairwise SSAP scores (> 90).

40-44. Endonucleases

Seven endonuclease families are represented in the current dataset. The *Fok*I family also belongs to the HTH group and has already been described. Figure 8b-f displays MolScript diagrams for representative structures of all the families, viewed parallel and perpendicular to the DNA axis. *Eco*RV, *Pvu*II, *Eco*RI and BamHI (1rva, 1piv, 1eri and 1bhm, respectively) are type II restriction endonucleases that recognize DNA sites of 6 bp in length and cleave the phosphate backbone at precise positions within the target sequence. Although there is little sequence similarity between the four protein types, their U-shaped homodimeric structures display some very common features [57,84-87].

The subunits of *Pvu*II (about 140 residues per subunit) and EcoRV (approximately 240 residues per subunit) may be divided into three segments: the amino-terminal dimerization region, the core catalytic region and the carboxy-terminal DNA-recognition region (Figure 8b,c). The catalytic regions of both comprise a five- or six-stranded mixed parallel/antiparallel β sheet (colored blue), which forms part of the cavity base. Most of the DNA-recognition segments extend from the carboxy-terminal end of the catalytic region (red). In *Pvu*II, the region comprises two parallel α helices and in *Eco*RV, a mixture of α helices and β strands. Both proteins approach the minor groove, and the DNA-recognition regions reach around the side of the DNA to contact bases in the major groove using a pair of loops. The dimerization regions of the two proteins are very different (colored green) and complete the base of the cavity [84,85].

The catalytic (or core) regions of endonucleases *Eco*RI (about 250 residues per monomer) and *Bam*HI (about 200 residues per monomer) also consist of five-stranded

parallel/antiparallel β sheets (Figure 8d,e). The positioning of the sheets is different from *Eco*RV and *Pvu*II, and they form the sides of the cavities. Included in the core region of both proteins are two α helices that pack against their counterparts in the other subunit to form a four-helix bundle at the base of the cavity. *Eco*RI and *Bam*HI both approach the DNA and make most of the base contacts from the major groove, although the method of sequence recognition greatly differ. *Eco*RI uses an extra set of interstrand loops and strands that follow the major groove towards the outer edges of the target sequence from the center (green in Figure 8d). *Bam*HI lacks these extra regions and uses the amino-terminal end of the helical bundle for binding [87,88].

44. Endonuclease V

This protein (for example 1vas) catalyzes the first step in the pyrimidine-specific base-excision repair pathway. In contrast to the type II enzymes described above, endonuclease V functions as a monomer (about 130 residues) whose structure comprises a four-helix bundle arranged to form a concave surface in which the DNA is bound (Figure 8f). Binding is centered on a damaged pyrimidine dimer; most of the interactions are to the DNA backbone, and the only base contacts are made to the central adenine which is flipped out of the DNA helix into a cavity on the protein surface.

45. DNase I

Function. DNase I is an endonuclease that degrades double-stranded DNA in a non-specific but sequence-dependent manner. Its function is dependent on the presence of divalent cations such as Ca^{2+}, Mg^{2+} and Mn^{2+}.

Structure. DNase I is an α,β protein with two six-stranded β-pleated sheets packed against each other forming the core of a 'sandwich'-type structure. The two predominantly antiparallel β sheets are flanked by three longer α helices and extensive loop regions.

Binding. DNase I binds in the minor groove of the DNA duplex with an exposed loop region forming contacts in and along both sides of the minor groove and extending over a total of 6 bp (Figure 8g). As a consequence of DNase I binding, the minor groove opens by about 3 Å and the duplex bends towards the major groove by about 20°.

46. DNA mismatch endonuclease

Function. In *E. coli*, the enzyme recognizes a TG mismatched base pair, generated after spontaneous deamination of methylated cytosines, and cleaves the phosphate backbone on the 5′ side of the thymine.

Structure. The protein contains three helices surrounding a β sheet, with one other helix used to intercalate the DNA.

Binding. Three aromatic residues from one helix intercalate into the major groove of the DNA to strikingly deform the base pair stacking (Figure 8h).

47-50. Polymerase group

Polymerases must provide sequence-independent interactions with their DNA substrate, yet retain the specificity to distinguish correctly paired bases from mismatches. DNA polymerases synthesize DNA strands by catalyzing the stepwise addition of a deoxyribonuleotide to the 3′-OH end of a polynucleotide chain that is paired to a second, template stand. Four polymerases have been classified: Pol β, Pol I, Pol T7 and Pol RT (reverse transcriptase).

47. DNA polymerase β (pol β); 48. DNA polymerase I (pol I); 49. DNA polymerase T7 (pol T7)

Pol β (Figure 8i) and Pol I (Figure 8j) have three structural domains that perform three separate functions, not only polymerizing the DNA but editing and repairing it by 3′-5′- and 5′-3′-exonuclease activity respectively. T7 DNA polymerase (Figure 8k) possesses no 5′-3′-exonuclease activity. For Pol I and T7, the larger carboxy-terminal domain has both the polymerase and 3′-5′-exonuclease activity with an α+β structure that can be likened to that of a right hand. A large cleft formed from a six-stranded antiparallel β sheet surrounded by α helices forms the 'palm' and binds the DNA minor groove along with the 'thumb' region (Figure 8j,k). Extensive sequence-independent interactions exist in the minor groove. The major groove, with its sequence-specific pattern of hydrogen-bond donors and acceptors, which form the primary means of recognition for many sequence-specific DNA-binding proteins, does not contact the protein and is solvent-accessible.

The smaller amino-terminal of Pol I has 5′-3′-exonuclease activity. It is folded into an αβ structure with a mixed β sheet of five strands.

50. HIV reverse transcriptase

Function. Reverse transcriptases have two enzymatic activities: a DNA polymerase that can copy either DNA or RNA templates and an RNase H. The two crystal structures of HIV reverse transcriptase which have been solved are only of the polymerase region.

Structure. HIV-1 reverse transcriptase (Figure 8l) is a heterodimer consisting of p66 (about 550 residues) and p51 (about 430 residues), two subunits of α helices and β strands which share a common amino terminus. The p51 subunit corresponds to the polymerase domain of the p66 subunit. The carboxy terminus of p66 forms the RNase H domain.

Binding. Loops and helices of p66 make extensive interactions with the DNA. P51 also binds but its interactions are mainly at the protein dimer interface with p66.

51. Uracil-DNA glycosylase

Function. Any uracil bases in DNA, a result of either misincorporation or deamination of cytosine, are removed by uracil-DNA glycosylase (UDG).

Structure. UDG is 225 residues long and contains a central four-stranded β-sheet region partly surrounded by eight α helices (Figure 8m).

Binding. Damaged DNA binds to UDG near the carboxy-terminal end of its central four-stranded β sheet. Conserved UDG residues in loop regions contact the DNA, with the loop between sheet 4 and helix 8 inserting into the DNA minor groove. A few contacts with the DNA backbone are made by two helices.

52. 3-Methyladenine DNA glycosylase

Function. DNA N-glycosylases are base excision-repair proteins that locate and cleave damaged bases from DNA as the first step in restoring the sequence.

Structure. The protein is 216 residues in length and is composed mainly of β strands (Figure 8n).

Binding. The enzyme intercalates into the minor groove of DNA using two β strands, causing the damaged base to flip into the enzyme active site for base excision.

53. Homing endonuclease family

Function. Homing endonucleases are a diverse collection of proteins that are encoded by genes with mobile, self-splicing introns. These enzymes promote the movement of the DNA sequences that encode them from one chromosome location to another; they do this by making a site-specific double-strand break at a target site in an allele that lacks the corresponding mobile intron.

Structure. The protein binds DNA as a dimer and displays mixed αβ topology (Figure 8o). Each monomer contains three antiparallel β sheets flanked by two long α helices, and a long carboxy-terminal tail that extends around the surface of the second subunit in the dimer and is stabilized by two bound zinc ions 15 Å apart.

Binding. The zinc-binding motifs are critical primarily for structural stabilization of the protein core and are not involved in DNA binding. The primary sequence-specific contacts made to homing-site DNA are from residues in the second β sheet of each enzyme monomer which contact the major groove of each half-site. Additional contacts are made in the center of the complex within the minor groove and with several phosphate groups in the cleavage site.

54. Topoisomerase I

Function. Topoisomerases I promote the relaxation of DNA superhelical tension by introducing a transient single-stranded break in duplex DNA and are vital for the processes of DNA replication, transcription and recombination.

Structure. No crystal structure has been solved for the whole protein - only for the central core and the carboxy-terminal domains (592 residues; see Figure 8p). The central core domain is connected to the carboxy-terminal domain by a linker. This linker assumes a coiled-coil configuration and protrudes away from the remainder of the enzyme.

Binding. The enzyme completely surrounds the DNA, contacting the backbone with loops and a β sheet binds in the major groove.

Acknowledgements

N.M.L. is supported by a BBSRC special studentship and S.E.A. by the US Department of Energy. This is a publication from the BBSRC Bloomsbury Centre for Structural Biology [http://www.biochem.ucl.ac.uk/BCSB].

References

1. Harrison SC: **A structural taxonomy of DNA-binding domains.** *Nature* 1991, **353:**715-719.
2. Luisi BF: **DNA-protein interaction at high resolution.** In *DNA-Protein Structural Interactions*. Edited by Lilley DMJ. New York: Oxford University Press, 1995, 1-48.
3. Frishman D, Mewes H-W: **PEDANTic genome analysis.** *Trends Genet* 1997, **13:**415-416.
4. Bernstein FC, Koetzler TF, Williams GJB, Meyer EF, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based achival file for macromolecular structures.** *J Mol Biol* 1977, **112:**535-542.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
6. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin, Demeny T, Hsieh S-H, Srinivasan AR, Schneider B: **The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids.** *J Biophys* 1992, **63:**751-759.
7. Sayle RA, Milner-White EJ: **RasMol – Biomolecular graphics for all.** *Trends Biochem Sci* 1995, **20:**374-376.
8. Orengo CA, Taylor WR: **SSAP: sequential structure alignment program for protein structure comparison.** *Methods Enzymol* 1996, **266:**617-635.
9. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequences of two proteins**. *J Mol Biol* 1970, **48:**443-453.
10. Orengo CA, Flores TP, Taylor WR, Thornton JM: **Identification and classification of protein fold families**. *Protein Eng* 1993, **6:**485-500.
11. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **24:**4876-4882.
12. Brennan RG, Matthews BW: **The HTH DNA-binding motif.** *J Biol Chem* 1989, **264:**1903-1906.
13. Harrison SC, Aggarwal AK: **DNA recognition by proteins with the helix-turn-helix motif**. *Annu Rev Biochem* 1990, **59:**933-969.
14. Pabo CO, Sauer, RT: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, **61:**1053-1095.
15. Steitz TA: *Protein-nucleic acid interactions.* Cambridge: Cambridge University Press, 1993.
16. Bacon D, Anderson, WF: **A fast algorithm for rendering space-filling molecule pictures.** *J Mol Graph* 1988, **6:**219-220.
17. Kraulis PJ: **MolScript – a program to produce both detailed and schematic plots of protein structures.** *J Appl Cryst* 1991, **24:**946-950.
18. Feng J-A, Johnson, RC, Dickerson RE: **Hin recombinase bound to**

DNA: the origin of specificity in major and minor groove interactions. *Science* 1994, **263:**348-355.
19. Suzuki M, Yagi N, Gerstein MB: **DNA recognition and superstructure formation by leix-turn-helix proteins.** *Protein Eng* 1995, **8:**329-338.
20. Lim WA, Hodel A, Sauer RT, Richards FM: **The crystal structure of a mutant protein with altered but improved hydrophobic core packing**. *Proc Natl Acad Sci USA* 1994, **91:**423-427.
21. Schumacher MA, Glasfeld A, Zalkin H Brennan RG: **The X-ray structure of the PurR-guanine-PurF operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated hydrogen bond to corepressor specificity and binding affinity**. *J Biol Chem* 1997, **272:**22648-22653.
22. Suzuki M, Brenner SE, Gerstein MB, Yagi, N: **DNA recognition code of transcription factors.** *Protein Eng* 1995, **8:**319-328.
23. Lawson CL, Carey J: **Tandem binding in crystals of a Trp repressor operator half-site complex.** *Nature* 1993, **366:**178-182.
24. Luisi BF: **DNA-transcription – zinc standard for economy.** *Nature* 1992, **356:**379-380.
25. MacKay JP, Crossely M: **Zinc fingers are sticking together.** *Trends Biochem Sci* 1998, **23:**1-4.
26. Jacobs GH: **Determination of the base recognition positions of zinc finger from sequence-analysis.** *EMBO J* 1992, **11:**4507-4517.
27. Pavletich NP, Pabo CO: **Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1Å.** *Science* 1991, **252:**809-817.
28. Suzuki M, Gerstein MB, Yagi, N: **Stereochemical basis of DNA recognition by Zn fingers.** *Nucleic Acids Res* 1994, **22:**3397-3405.
29. Fairall L, Schwabe JWR, Chapman L, Finch JT, Rhodes D: **The crystal structure of a two zinc finger peptide reveals an extension to the rules of zinc finger DNA recognition.** *Nature* 1993, **366:**483-487.
30. Choo Y, Klug A: **Selection of DNA-binding sites for zinc fingers using rationally randomized DNA reveals coded interactions.** *Proc Natl Acad Sci USA* 1994, **91:**11168-11172.
31. Pavletich NP, Pabo CO: **Crystal structure of a 5-finger GLI-DNA complex – new perspectives on zinc fingers.** *Science* 1993, **261:**1701-1707.
32. Freedman LP, Luisi BF: **On the mechanism of DNA-binding by nuclear hormone receptors – a structural and functional perspective.** *J Cell Biochem* 1993, **51:**140-150.
33. Schwabe JW, Chapman L, Finch JT, Rhodes D: **The crystal structure of the estrogen-receptor DNA-binding domain bound to DNA – how receptors discriminate between their response elements.** *Cell* 1993, **75:**567-578.
34. Cho Y, Gorina S, Jeffrey PD, Pavletich NP: **Crystal structure of a p53 tumor supressor-DNA complex: understanding tumorgenic mutations.** *Science* 1994, **265:**346-355.
35. Marmorstein R, Carey M, Ptashne M, Harrison SC: **DNA recognition by GAL4: structure of a protein-DNA complex.** *Nature* 1992, **356:**408-414.
36. Ellenberger TE, Brandl CJ, Struhl K, Harrison SC: **The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: crystal structure of the protein-DNA complex.** *Cell* 1992, **71:**1223-1237.
37. Ferre d'Amare AR, Prendergast GC, Ziff EB, Burley SK: **Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain.** *Nature* 1993, **363:**38-45.
38. Phillips SE: **Built by association – structure and function of helix-loop-helix DNA-binding proteins.** *Structure* 1994, **2:**1-4.
39. Rupert PB, Daughdrill GW, Bowerman B, Matthews BW: **A new DNA-binding motif in the Skn-1 binding domain-DNA complex.** *Nat Struct Biol* 1998, **5:**484-491.
40. Tan S, Richmond TJ: **Crystal structure of the yeast Matα2/MCM1/DNA ternary complex.** *Nature* 1998, **391:**660-666.
41. Hegde RS, Grossman SR, Laimins, Sigler PB: **Crystal structure at 1.7Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target.** *Nature* 1992, **359:**505-512.
42. Bochkarev A, Bochkareva E, Edwards AM, Frappier L: **The 2.2Å structure of a permanganate sensitive DNA site bound by the Epstein-Barr virus origin-binding protein, Ebna 1.** *J Mol Biol* 1998, **284:**1273-1278.
43. Luger K, Mader AW, Richmod RK, Sargent DF, Richmod TJ: **Crystal structure of the nucleosome core particle at 2.8Å resolution.** *Nature* 1997, **389:**251-260.
44. Murphy FV, Sweet RM, Churchill MEA: **The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition.** *EMBO J* 1999, **18:**6610-6618.
45. Guo F, Gopaul DN, van Dyke GD: **Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse.** *Nature* 1997, **389:**40-46.
46. Kim Y, Geiger JH, Hahn S, Sigler PB: **Crystal structure of yeast TBP/TATA-box complex.** *Nature* 1993, **365:**512-520.
47. Burley SK: **The TATA box binding protein.** *Curr Opin Struct Biol* 1996, **6:**69-75.
48. Somers WS, Phillips SEV: **Crystal structure of the Met repressor-operator complex at 2.8Å resolution reveals DNA recognition by beta strands.** *Nature* 1992, **359:**387-391.
49. Raumann BE, Rould MA, Pabo CO, Sauer RT**: DNA recognition by β-sheets in the Arc repressor-operator crystal structure.** *Nature* 1994, **367:**754-757.
50. Kamada K, Horiuchi T, Ohsumi K, Shimamoto N, Morikawa K: **Structure of a replication-terminator protein complexed with DNA.** *Nature* 1996, **383:**598-603.
51. Muller CW, Herrmann BG: **Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor.** *Nature* 1997, **389:**884-888.
52. Rice PA, Yang S-W, Mizuchi K, Nash HA: **Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn.** *Cell* 1996, **87:**1295-1306.
53. Robinson H, Gao YG, McCray BS, Edmondson SP, Shriver JW, Wang AHJ: **The hyperthermophile chromosomal protein Sac7D sharply kinks DNA.** *Nature* 1998, **392:**202-205.
54. Ghosh G, Van Duyne G, Ghosh S, Sigler PB: **Structure of NF-κB p50 homodimer bound to a kappa B site.** *Nature* 1995, **373:**303-310.
55. Chen X, Vinkemeier U, Zhao Y, Jeruzalmi D, Darnell JE Jr., Kuriyan J: **Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA.** *Cell* 1998, **93:**827-839.
56. Jones S, van Heyningen P, Berman HM, Thornton JM: **Protein-DNA interactions: a structural analysis.** *J Mol Biol* 1999, **287:**877-896.
57. Aggarwal AK: **Structure and function of restriction endonucleases.** *Curr Opin Struct Biol* 1995, **5:**11-19.
58. Attwood TK, Croning MDR, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley J, Wright W: **PRINTS-S: the database formerly known as PRINTS.** *Nucleic Acids Res* 2000, **28:**225-227.
59. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM: **PDBsum: a Web-based database of summaries and analyses of all PDB structures.** *Trends Biochem Sci* 1997, **22:**488-490.
60. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH - a hierarchic classification of protein domain structures.** *Structure.* 1997, 5:1093-1108.
61. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of proteins structures.** *J Appl Crystallogr* 1993, **26:**283-291.
62. Hutchinson EG, Thornton JM: **PROMOTIF – a program to identify and analyze structural motifs in proteins.** *Protein Sci* 1996, **5:**212-220.
63. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247:**536-540.
64. Vriend G: **WHAT IF: a molecular modeling an drug design program.** *J Mol Graph* 1990, **8:**52-56.
65. Holm L, Sander C: **Mapping the protein universe.** *Science* 1996, **273:**595-602.
66. Beamer LJ, Pabo CO: **Refined 1.8Å crystal structure of the lambda repressor-operator complex.** *J Mol Biol* 1992, **227:**177-196.
67. Schumacher MA, Choi KY, Zalkin H, Brennan RG: **Crystal structure of LacI family member, Purr, bound to DNA: minor groove binding by alpha helices.** *Science* 1994, **266:**763-770.
68. Kercher MA, Lu P, Lewis M: **Lac repressor-operator complex.** *Curr Opin Struct Biol* 1997, **7:**76-85.
69. Pace HC, Kercher MA, Lu P, Markiewicz P, Miller JH, Chang G, Lewis M**: Lac repressor genetic map in real space.** *Trends Biochem Sci* 1997, **22:**334-9.

70. Wah DA, Hirsch JA, Dorner LF, Schildkraut I, Aggarwal AK: **Structure of the multimodular endonuclease FokI bound to DNA.** *Nature* 1997, **388:**97-100.
71. Yang W, Steitz TA: **Crystal structure of the site-specific recombinase gamma-delta resolvase complexed with a 34-bp cleavage site.** *Cell* 1995, **82:**193-207.
72. Konig P, Giraldo R, Chapman L, Rhodes D: **The crystal structure of the DNA-binding domain of yeast Rap1 in complex with telomeric DNA.** *Cell* 1996, **85:**125-136.
73. Xu W, Rould MA, Jun S, Desplan C, Pabo CO: **Crystal structure of a paired domain-DNA complex at 2.5Å resolution reveals structural basis for Pax developmental mutations.** *Cell* 1995, **80:**639-650.
74. Van Pouderoyen G, Ketting RF, Perrakis A, Plasterk RHA, Sixma TK: **Crystal structure of the specific DNA-binding domain of Tc3 transposase of *C. elegans* in complex with transposon DNA.** *EMBO J* 1997, **16:**6044-6054.
75. Schultz, SC, Shields, GC, Steitz, TA**: Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees**. *Science* 1991, **253:**1001-1007.
76. Suzuki M, Gerstein, M: **Binding geometry of alpha helices that recognize DNA.** *Proteins* 1995, **23:**525-535.
77. Branden C, Tooze J: *Introduction to Protein Structure.* New York; Garland Publishing, 1991.
78. Choo Y, Klug A: **Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage.** *Proc Natl Acad Sci USA* 1994, **91:**11163-11167.
79. Alber T: **Structure of the leucine zipper.** *Curr Opin Genet Dev* 1992, **2:**205-10.
80. Rice PA: **Making DNA do a U-turn: IHF and related proteins.** *Curr Opin Struct Biol* 1997, **7:**86-93.
81. Chytil M, Verdine GL: **The Rel family of eukaryotic transcription factors.** *Curr Opin Struct Biol* 1996, **6:**91-100.
82. Klimasauskas S, Kumar S, Roberts RJ, Cheng X: **Hha1 methyltransferase flips its target base out the DNA helix.** *Cell* 1994, **76:**357-369.
83. Cheng X: **DNA modification by methyltransferases.** *Curr Opin Struct Biol* 1995, **5:**4-10.
84. Kostrewa D, Winkler FK: **Mg^{2+} binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2Å resolution.** *Biochemistry* 1995, **34:**683-696.
85. Cheng X, Balendiran K, Schildkraut I, Anderson JE: **Structure of PvuII endonuclease with cognate DNA.** *EMBO J* 1994, **13:**3927-3935.
86. Kumar S, Duan Y, Kollman PA, Rosenberg JM: **Molecular dynamics simulations suggest that the Eco RI kink is an example of molecular strain.** *J Biomol Struct Dyn* 1994, **12:**487-525.
87. Newman M, Strzelecka T, Dorner LF, Schildkraut I, Aggarwal AK: **Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding.** *Science* 1995, **269:**656-663.
88. Kim Y, Grable JC, Love R, Greene P, Rosenberg JM: **Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing.** *Science* 1990, 249:1307-1309.