

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, NJ 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):reviews001.1–001.10

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

© GenomeBiology.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 the PEDANT database [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 $\beta\beta\alpha$ 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.

*The full version of this article is available online at <http://genomebiology.com> and includes figures and text that are not reproduced in this printed version.

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 about 240 protein-DNA complexes (Table 1). 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 (available online only; see <http://genomebiology.com/2000/1/1/reviews/001>).

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 features 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 (see Table 2, online). 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 endonucleases *FokI* (PDB entry 1fok), $\gamma\delta$ -resolvase (1gdt), Hin recombinase (1hcr), Tc3 transposase (1tc3) and Cre recombinase (1crx; these proteins are listed under the HTH group in Table 2 online 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]. This 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 $\beta\beta\alpha$ zinc fingers, however, a chain contains several binding domains; in such cases, therefore, the subunits were separated into the appropriate segments, which are listed online in Table 2. In this review, structures are identified by the standard four-digit PDB code (for example, 1aay). Where a protein subunit is specified, the corresponding chain identity in the PDB file

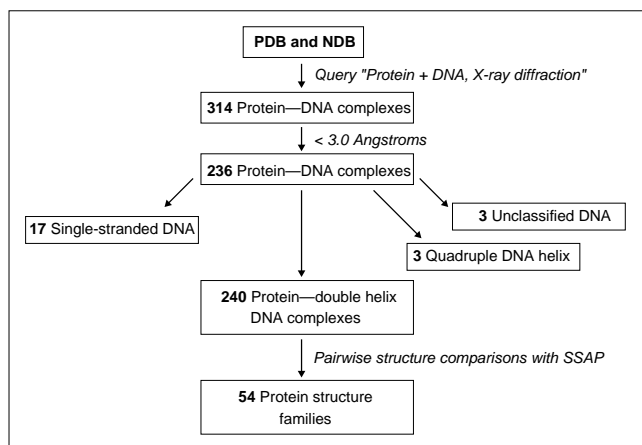
Table 1

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.

Protein group	Number of families in group	Number of proteins (PDB files) in group			Total
		Prokaryote	Eukaryote	Viral	
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'.

**Box 1**

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

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 (online) lists all the protein-DNA complex structures in the dataset and their classifications. 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. This review is available online [<http://genomebiology.com/2000/1/1/reviews/001>], together with summaries of the individual families, and additional tables and figures.

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 1), 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 online.

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; see figure online). 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

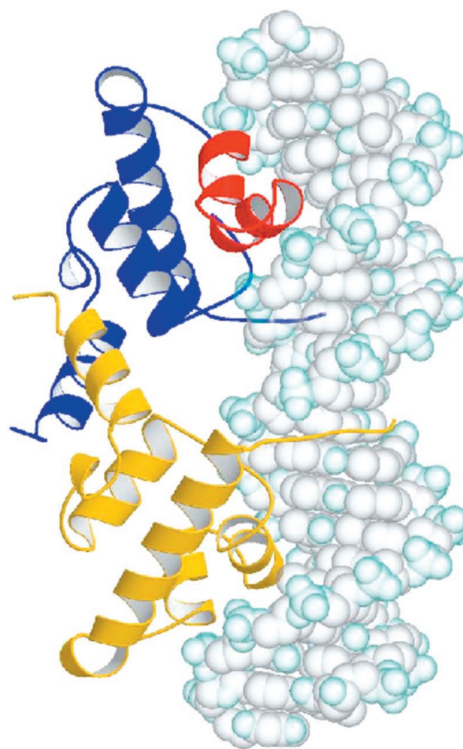


Figure 1
Group I, HTH proteins: the Cro and Repressor family (1lmb). The DNA-binding motif is red. The protein binds as a dimer; one monomer is colored blue and the other yellow. The DNA is shown as a space-filling model. Additional examples of Group I are shown online.

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, *FokI* 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 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 *Mat α -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, $\gamma\delta$ -resolvase, *FokI* 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 , 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 online, 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 2). The name 'zinc-coordinating' will be used as the generic term for all proteins with zinc ions in the DNA-binding motif.

The $\beta\beta\alpha$ zinc-finger family

The $\beta\beta\alpha$ 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 2). 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. 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 heterodimers 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. 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

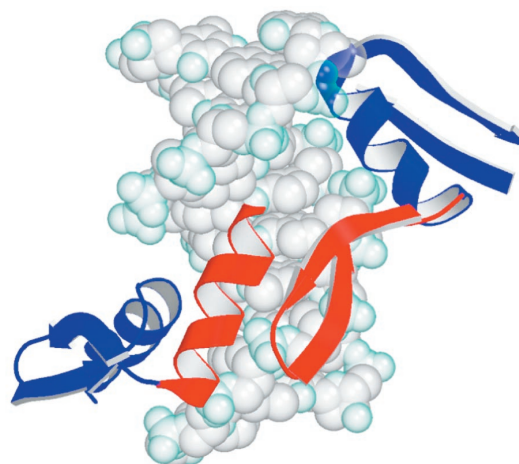


Figure 2
Group II, zinc-coordinating proteins: The $\beta\beta\alpha$ zinc finger motif (1aay). Colors are used as in Figure 1; additional examples of Group II are illustrated online.

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 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. 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 3) and helix-loop-helix proteins, 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 3, 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. 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

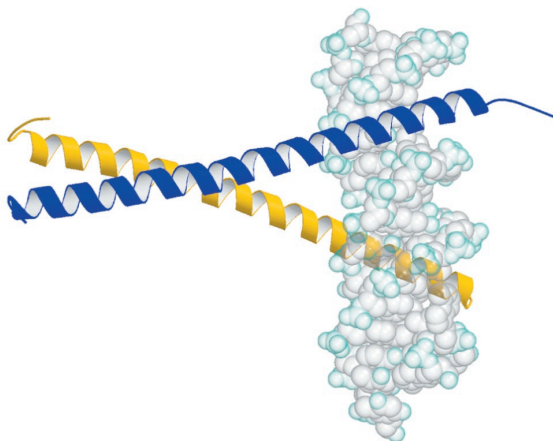


Figure 3
Group III, zipper-type proteins: the leucine zipper motif (2dge). Colors are used as in Figure 1; additional examples of Group III are illustrated online.

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. Skn-1 (1skn) and MADS (see Figure 4 for the MADS box, 1mmn) are transcription regulatory regions in eukaryotic proteins, papillomavirus-1 E2 (2bop) and EBNA1 (1b3t) are viral transcription regulators and replication initiators, histones (1aoi) and high-mobility group (HMG) proteins (1qrv) are architectural proteins for DNA packaging, and Cre (1crx) is a site-specific recombinase. Although the protein structures are very different, all use α helices 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 is monomeric with a compact four-helix unit (see Figure online); 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 4), an antiparallel β sheet and an adjacent coiled-coil provide the dimerization interface. The α helices on the opposite face of

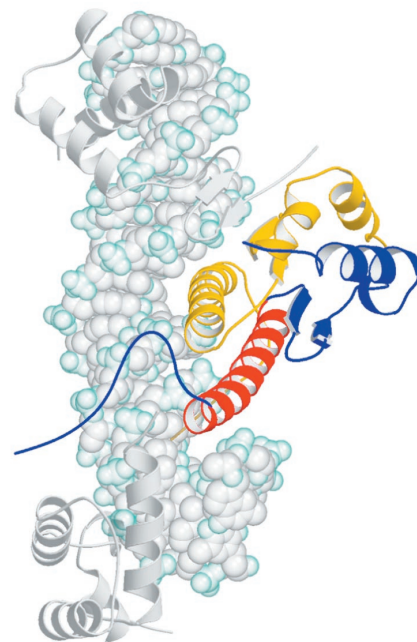


Figure 4
Group IV, 'other α helix' proteins: the MADS box (1mmn). Colors are used as in Figure 1; additional examples of Group IV are illustrated online.

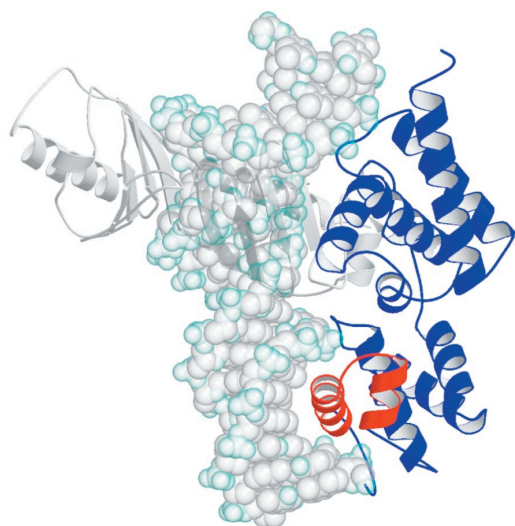


Figure 5
Group V, the β -sheet proteins: the TATA box-binding family (1ytb). Colors are used as in Figure 1; additional examples of Group V are illustrated online.

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. 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 groove symmetrically. As is apparent from their structures, 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 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. 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 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. Six protein families of very diverse function are represented: the MetJ repressor (1cma; Figure 6), Arc repressor (1bdt) and T-domain families (1xbr) constitute DNA-binding regions of transcriptional regulators; the integration host factor (1ihf) and the hyperthermophile chromosomal proteins (1azp) act as scaffolds to dictate the DNA structure for formation of high-order protein-DNA complexes; and the Tus protein (1ecr) terminates DNA replication by helicases (all are illustrated online). Although the overall structures of the proteins are different, there are common themes in the use of the β strands.

The MetJ and Arc repressor are both dimers with very similar modes of binding (see Figure 6). 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. 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. 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. 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 is a conserved amino-terminal domain of transcriptional regulators involved in cellular

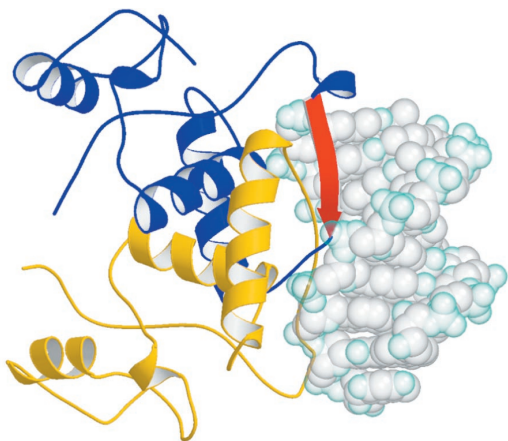


Figure 6
Group VI, the β hairpin/ribbon proteins: Met repressor (1cma). Colors are used as in Figure 1; additional examples of Group VI are illustrated online.

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 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]. Illustrations of members of this group are available with the online version of this article.

Group VIII: the enzymes

The enzyme group completes the classification of the dataset. 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 the 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. As described in the online version of this article, 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 *HhaI* methyltransferases and endonucleases, a single change in the target sequence can lead to over a million-fold reduction in binding affinity. Proteins are thought to derive their specificity from both read-out of the base sequence and the catalytic action on the DNA, as in endonucleases *BamHI* (3bam) and *EcoRI* (1qps), or even primarily from the catalytic process, as in endonuclease *EcoRV* (1rva) [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 dominate this group of 16 families.

A protein-DNA complex website

As well as the online version of this article, 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 with this

review online, 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 be 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 yet well understood.

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 $\beta\beta\alpha$ 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.

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

- Harrison SC: **A structural taxonomy of DNA-binding domains.** *Nature* 1991, **353**:715-719.
- 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.
- Frishman D, Mewes H-W: **PEDANTic genome analysis.** *Trends Genet* 1997, **13**:415-416.
- 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 archival file for macromolecular structures.** *J Mol Biol* 1977, **112**:535-542.
- 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.
- Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin, Demeny T, Hsieh S-H, Srinivasan AR, Schneider B: **Relational database of three-dimensional structures of nuclei acids.** *J Biophys* 1992, **63**:751-759.
- Sayle RA, Milner-White EJ: **RasMol - Biomolecular graphics for all.** *Trends Biochem Sci* 1995, **20**:374-376.
- Orengo CA, Taylor WR: **SSAP: sequential structure alignment program for protein structure comparison.** *Meth Enzymol* 1996, **266**:617-635.
- 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.
- Orengo CA, Flores TP, Taylor WR, Thornton JM: **Identification and classification of protein fold families.** *Prot Eng* 1993, **6**:485-500.
- 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. Aggarwal AK, Harrison SC: **DNA recognition by proteins with the HTH 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 super-structure formation by HTH proteins.** *Prot 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.** *Prot 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, Crossley 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 2 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 JWR, 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 suppressor-DNA complex: understanding tumorigenic 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 SEV: **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 Mat α 2/MCM1/DNA ternary complex.** *Nature* 1998, **391**:660-666.
41. Hedge 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 AWW, Richmond RK, Sargent DF, Richmond 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, S.Hahn S, P.B.Sigler PB: **Crystal structure of YTBP/TATA-box complex.** *Nature* 1993, **365**:512-514.
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**:1295-1307.
53. Robinson H, Gao YG, McCray BS, Edmondson SP, Shriver JW, Wang AH: **The hyperthermophile chromosomal protein SAC7D sharply kinks DNA.** *Nature* 1998, **392**:202-205.
54. Ghosh G, Van Duyn G, Ghosh S, Sigler PB: **Structure of NF- κ B 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.** *Prot 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.