

Protein–DNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity

Nicholas M. Luscombe^{1*} and Janet M. Thornton^{1,2}

¹*Biomolecular Structures and Modelling Unit
Department of Biochemistry and Molecular Biology
University College
Gower Street
London WC1E 6BT, UK*

²*Department of Crystallography
Birkbeck College, Malet Street
London WC1E 7HX, UK*

We investigate the conservation of amino acid residue sequences in 21 DNA-binding protein families and study the effects that mutations have on DNA-sequence recognition. The observations are best understood by assigning each protein family to one of three classes: (i) non-specific, where binding is independent of DNA sequence; (ii) highly specific, where binding is specific and all members of the family target the same DNA sequence; and (iii) multi-specific, where binding is also specific, but individual family members target different DNA sequences. Overall, protein residues in contact with the DNA are better conserved than the rest of the protein surface, but there is a complex underlying trend of conservation for individual residue positions. Amino acid residues that interact with the DNA backbone are well conserved across all protein families and provide a core of stabilising contacts for homologous protein–DNA complexes. In contrast, amino acid residues that interact with DNA bases have variable levels of conservation depending on the family classification. In non-specific families, base-contacting residues are well conserved and interactions are always found in the minor groove where there is little discrimination between base types. In highly specific families, base-contacting residues are highly conserved and allow member proteins to recognise the same target sequence. In multi-specific families, base-contacting residues undergo frequent mutations and enable different proteins to recognise distinct target sequences. Finally, we report that interactions with bases in the target sequence often follow (though not always) a universal code of amino acid–base recognition and the effects of amino acid mutations can be most easily understood for these interactions

© 2002 Elsevier Science Ltd. All rights reserved

Keywords: bioinformatics; structural biology; protein–DNA interactions; transcription factors; sequence conservation

*Corresponding author

Introduction

DNA-binding proteins have a central role in all aspects of the genetic activity within an organism, such as transcription, packaging, rearrangement, replication and repair. It is therefore of great importance to understand the nature of inter-

actions between proteins and DNA, as they form the basis of our knowledge 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. These structures, especially those of complexes with DNA, have provided valuable insight into the molecular basis of binding, including the way that particular DNA sequences are recognised.

Most studies have examined the interactions found within individual structures. However, in addition, there have been numerous surveys in search for common principles of binding that apply across most, or all protein–DNA complexes.¹ Such analyses have been conducted at many levels, including atomic contacts between amino acid residues and bases,^{2–4} the usage of secondary structural elements and small structural

Present addresses: N. M. Luscombe, Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, P.O. Box 208114, New Haven, CT 06520-8114, USA; J. M. Thornton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.

Abbreviations used: IHF, integration host factor; HMM, hidden Markov model; PDB, Protein Data Bank; RHR, Rel homology region; CAP, catabolite gene regulator.

E-mail address of the corresponding author: nicholas.luscombe@yale.edu

Table 1. The number of member proteins in a sequence family

Family	Representative PDB structure	PDB	SWISS-PROT	Total combined	Total non-redundant (<95% sequences ID)
(i) Non-specific					
Integration host factor (IHF)	1ihf	4 (9)	51	60	34
Polymerase Taq	1tau	16 (16)	27	43	25
Polymerase- β	1bpy	54 (54)	9	63	11
Dnase I	2dnj	2 (2)	9	11	9
(ii) Highly specific					
Pu.1 ETS domain	1pue	9 (12)	52	64	26
Prd paired domain	1pdn	1 (1)	35	36	14
Trp repressor	1trr	9 (22)	5	27	8
Loop-sheet-helix (LSH)	1tsr	2 (6)	22	28	13
Leucine zipper	2dgc	3 (6)	40	46	18
Papillomavirus-1 E2	2bop	2 (3)	73	76	69
TATA box-binding protein (TBP)	1ytb	5 (6)	36	42	14
T-domain	1xbr	2 (2)	32	34	27
Rel homology region (RHR)	1nfk	2 (4)	21	25	14
(iii) Multi-specific					
Homeodomain	1yrn	6 (13)	465	478	359
LacI	1wet	4 (8)	38	46	34
CAP	1ber	5 (9)	23	32	21
$\gamma\delta$ -resolvase	1gdt	6 (11)	25	36	21
C ₂ H ₂ -zinc finger	1aay	5 (21)	269	290	213
Hormone receptor	2nll	4 (10)	159	169	63
GAL4	1d66	4 (5)	34	39	22
Helix-loop-helix (HLH)	1hlo	5 (10)	62	72	29

Families are divided into three binding-classes: non-specific, highly specific and multi-specific. The first two columns provide the name of the family and the PDB code of the representative complex structure. The central columns provide the number of PDB entries, the number of SWISS-PROT entries and the total number of sequences in the family (PDB entries + SWISS-PROT entries). The final column gives the non-redundant total, which removes sequences with >95% sequence identity and leaves just one representative.

motifs,⁵⁻⁸ up to how the whole protein structure interacts with the DNA.⁹⁻¹¹ The overriding conclusion from all of these analyses is that there is no simple code relating amino acid sequence to the DNA sequence it binds. Given the wide variety of secondary and tertiary structural frameworks that proteins use for binding,^{11,12} combined with the complexities of atomic movements in the protein and DNA,^{9,10,13} it is now clear that detailed rules for DNA-sequence recognition is best understood within the context of individual protein families,¹⁴ but with strong underlying trends, such as the preference for arginine-guanine interactions.⁴

The current study combines the two elements, (1) the general rules for amino acid-base interactions and family specific analysis, and (2) to examine the conservation of protein-DNA interactions between homologous proteins. Previous studies have used sequence and structural data to explore conservation patterns for different types of interactions. These were based on the premise that in order to retain a protein function between homologues, residues important for the task should be well conserved. Investigations of binding sites

further benefit from the fact that the residues concerned are on the protein surface; as surface conservation tends to be lower than in the protein core, regions of high conservation should stand out.

Starting with large binding sites, Grishin & Phillips¹⁵ analysed the rates of mutations at the protein-protein interface in five oligomeric enzymes and suggested that subunit interfaces are conserved only slightly better than the rest of the protein. More recently, however, a study of six dimeric protein families by Valdar & Thornton¹⁶ challenged this result, by using a scoring system that tolerated substitutions between similar amino acid residues. They concluded that interface residues are almost always better conserved than randomly selected residues on the protein surface.

Turning to interfaces with smaller ligands, Lichtarge *et al.*¹⁷ examined binding sites of SH2 and SH3 domains. Since different domains are specific for different peptide sequences, the authors argued that a simple search for invariant residue positions was inappropriate. Instead, a technique called evolutionary tracing was used to recognise groups of amino acids that mutate

between functionally distinct proteins. Mapping these residue positions onto a representative structure showed a good correspondence with binding regions on the protein surface. A similar method detected receptor and effector-binding sites on G proteins.^{18,19}

So far, the only equivalent study for DNA-binding proteins has been by Lichtarge *et al.*,²⁰ who conducted an evolutionary trace analysis on one family of intracellular nuclear receptors (including hormone receptors). They examined the DNA-binding, effector-binding and dimerisation interfaces, and highlighted many of the protein residues involved in recognition of the base sequence.

Here, we present the first global analysis of the conservation of amino acid residue sequences in DNA-binding proteins. We inspect a total of 21 distinct protein families by combining data of over 150 complex structures from the Protein Data Bank (PDB), 1500 protein sequences from SWISS-PROT and associated target DNA sequences. The aims of the study are twofold. First is to see whether amino acid residues that interact with DNA are better conserved than the rest of the protein surface and the second is to assess the effect that amino acid mutations have on binding specificity. In doing so, we explore the molecular basis for DNA sequence recognition, and we consider the evolutionary relationship between related DNA-binding proteins.

Results

Summary of methods

Briefly, we: (i) compiled structural families of protein–DNA complexes from the PDB,^{21,22} (ii) expanded the families to include sequence homologues from SWISS-PROT;²³ (iii) produced multiple sequence alignments of each family; (iv) calculated a conservation score for each amino acid residue position in a family; and (v) identified amino acid residue positions that reside on the protein surface or interact with DNA.

The procedure allowed us to study the amino acid conservation of 21 DNA-binding protein families, containing nine to 359 members (Table 1). For the rest of the work, we refer to protein structures by their four-letter PDB code and we refer to sequences by their SWISS-PROT entry names. A detailed description of the methods is provided in Materials and Methods.

Three-classes of protein families

The results we present are best understood by classifying the protein families into one of three classes on the basis of their DNA-binding specificities. As summarised in Figure 1, the classes are (i) non-specific, (ii) highly specific and (iii) multi-specific. Non-specific families comprise proteins

that bind promiscuously and have no requirement for any specific base sequence. Highly specific families consist of proteins that bind DNA specifically, and all members of a family target a common base sequence. Multi-specific families also contain proteins that bind specifically, but different members bind distinct and different targets.

We define four non-specific, nine highly specific and eight multi-specific families (Table 1). Each family is assigned to a class according to published data on target DNA sequences. Biological targets are used where possible but, when unavailable, non-biological targets defined through binding assays are used. We note that a target-site definition does not preclude proteins from binding other DNA sequences. Although members of the integration host factor (IHF) and DNase I families display slight sequence-dependence, we consider them to be non-specific, as their binding-preferences are not strong enough to define clear target sequences.^{24,25}

Number of DNA-binding positions

The average length of a multiple alignment is 138 amino acid residue positions, including gaps. As we are concerned with amino acid residues that can potentially interact with the DNA, we consider only positions that are at the protein surface.

On average, families have 23 interacting positions, with numbers ranging from 14 and 47 for individual families (Table 2). These correspond to 6.8% and 63.3% of the total number of alignment positions in each family, and the proportion of amino acid residue positions in contact with the DNA is very variable between families

We observe an interesting trend when interactions are divided into those with DNA bases and those with the DNA backbone (Table 2). Ratios of backbone to base-contacting positions are high for non-specific families (mean ratio backbone-contacting to base-contacting positions = 4.4), indicating that many more protein residues interact with the DNA backbone than with bases. In contrast, ratios are lower for multi-specific (mean ratio = 3.2) and highly specific families (mean ratio = 3.5), demonstrating that there is greater emphasis towards interactions with bases in these proteins. If we consider only direct read-out of the DNA sequence, DNA backbone contacts mainly provide stability to the complex, while base contacts also contribute specificity.⁴ (Amino acid residues that interact with both are considered to be “base-contacting” as they potentially provide direct sequence recognition.) Thus, the ratios for sequence-specific proteins reflect the relative importance of interactions with DNA bases.

Overall conservation of protein sequences

Figures 2–4 show histograms of conservation scores (*c*) for amino acid residue positions in the

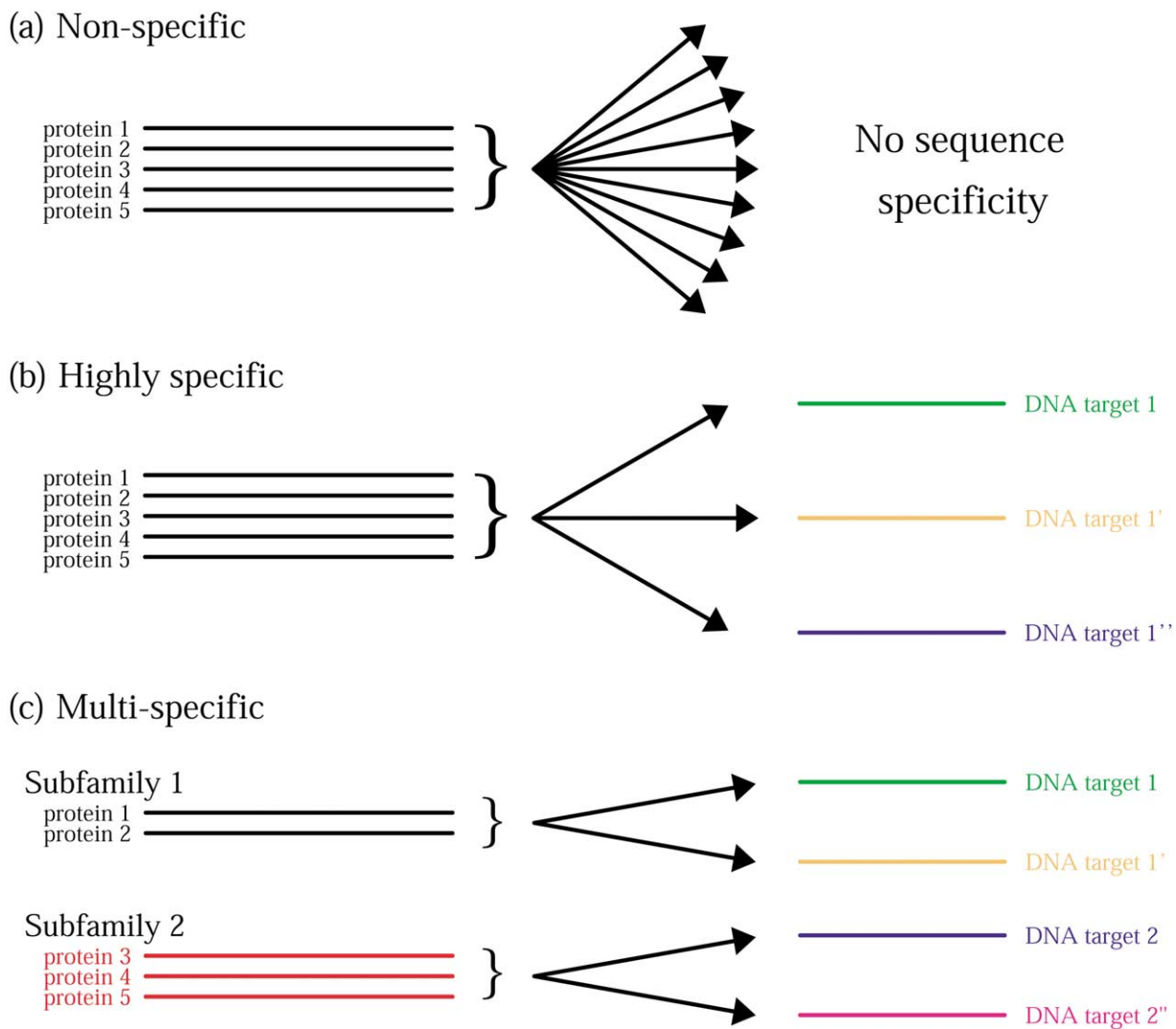


Figure 1. Schematic representing the mode of DNA-sequence recognition in the three binding classes. Lines on the left represent member of a sequence family, each line being a different protein. Lines on the right represent the target sequences, each line being a different DNA sequence. The interconnecting arrows show recognition of each base sequence by the proteins. (a) Non-specific families bind promiscuously and do not target particular DNA sequences. (b) Highly specific families contain proteins that are all specific for the same DNA sequence. Members may recognise several related sequences. (c) Multi-specific families contain proteins that are specific but individual members target different DNA sequences. However, members of the same subfamily (shown in red and black) target the same sequences.

21 families. Scores ranges from 0 (unconserved) to 100 (conserved) for each position, and we record the number of positions with a given score for each family. Histogram bars are coloured by whether the amino acid residue position interacts with a DNA base (black) or only the sugar-phosphate backbone (orange).

We start by looking at the overall conservation of the whole protein surface. Generally, the longer the alignment, the smoother the distribution, so alignments with more than 100 positions have continuous distributions and those with fewer positions have irregular, discontinuous distributions. Overall, non-specific and highly specific families have greater amino acid conservation than multi-specific families; this is reflected in the mean surface conservation scores, \bar{c}_{surf} (Table 3).

The variation in sequence conservation between the binding classes reflects two factors: (i) differences in family sizes; and (ii) differences in diversity of protein functions. First, non-specific and highly specific families have an average of 19.8 and 22.7 members each (Table 1), whereas multi-specific families contain an average of 87.9 members, including the highly populated C_2H_2 -zinc finger and homeodomain families. Furthermore, non-specific and highly specific families comprise proteins from an average of 14.5 and 8.7 organisms respectively, whereas multi-specific families contain members from 17.6 organisms. Second, a simple-minded survey of SWISS-PROT protein names shows that multi-specific families are represented by an average of 38 protein "functions". This is in contrast to the uniform functions of all

Table 2. The number of alignment positions that interact with the DNA

Family		Total	Backbone	Base	Ratio backbone/base
Non-specific					
IHF	1ihfA	28	20	8	2.5
Polymerase Taq	1tauA	28	22	6	2.7
Polymerase-b	1bpyA	47	40	7	5.7
Dnase I	2dnjA	20	17	3	5.7
Mean		30.8	24.8	6.0	4.4
Highly specific					
Pu1 ETS domain	1pueE	22	19	2	6.3
Prd paired domain	1pdnC	27	19	8	2.4
Trp repressor	1trrA	22	19	2	6.3
Loop-sheet-helix	1tsrB	17	14	2	4.7
Leucine Zipper	2dgcA	14	10	3	2.5
Papillomavirus-1 E2	2bopA	17	13	3	3.3
TBP	1ytbA	25	12	13	0.9
T-domain	1xbrA	22	18	4	4.5
RHR	1nfkA	24	12	12	1.0
Mean		21.1	15.1	6.0	3.5
Multi-specific					
Homeodomain	1yrnA	18	10	8	1.3
LacI	1wetA	23	17	6	2.8
CAP	1berA	19	14	5	2.8
gd-resolvase	1gdtA	25	16	9	1.8
C ₂ H ₂ -zinc finger	1aayA1	18	12	6	2.0
Hormone receptor	2nllA	34	26	8	3.3
GAL4	1d66A	13	10	3	3.0
HLH	1hloA	20	12	8	1.5
Mean		21.3	14.6	6.6	2.3
Overall mean		23.0	16.8	6.2	3.2

The total number of alignment positions that interact with the DNA are given, along with the number that interact with the DNA backbone, nucleotide bases and the ratio between the two.

non-specific and highly specific families except for the leucine zipper (eight functions) and Rel homology region (RHR) (six functions) families. Therefore, the genes for proteins in multi-specific families have duplicated and multiplied many more times during evolution to give much larger family sizes containing greater diversity of protein functions.

Overall conservation of interacting residue positions

Amino acid residues in contact with the DNA are generally better conserved than those that are not (Table 3). Mean conservation scores for interacting residue positions (\bar{c}_{int}) range from 50.6 to 97.0, and are higher than the mean surface scores

in all but the $\gamma\delta$ -resolvase family, which we discuss in greater detail below.

We find that the level of conservation of interacting residue positions is very dependent on the conservation of the entire protein sequence; the Pearson correlation coefficient between the mean scores for surface (\bar{c}_{surf}) and interacting positions (\bar{c}_{int}) is 0.85. Thus the conservation of DNA-binding residues reflects the extent of sequence divergence in different families and, as explained above, these are more conserved in highly specific and non-specific families than they are in multi-specific families (Table 3).

Although, on average, interacting residue positions are more conserved than non-interacting ones, there is in fact a complex underlying trend. Examination of the histograms in Figures 2–4

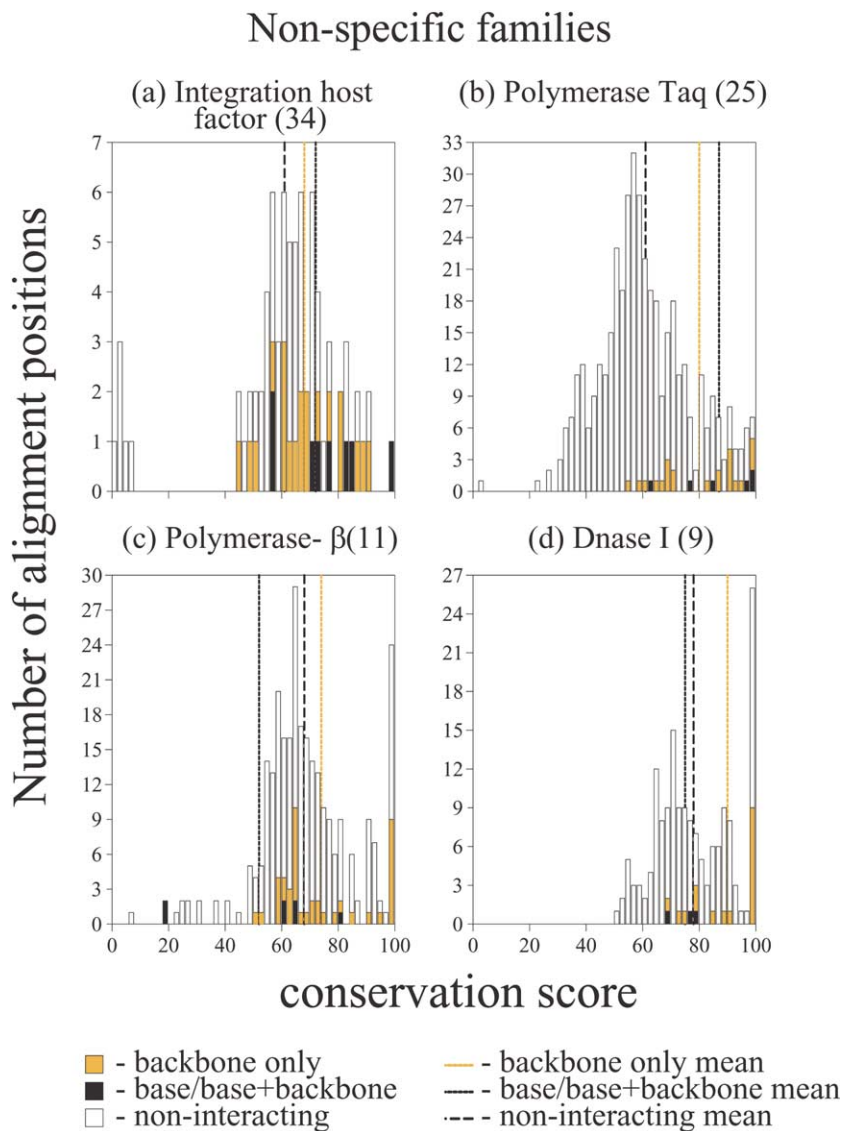


Figure 2. Conservation score distributions of surface alignment positions for non-specific families. Positions that interact with DNA bases are coloured black and those that interact with the backbone only are coloured orange. Broken lines mark the mean conservation scores: all surface positions, \bar{c}_{surf} (black long dashes); base-binding positions, \bar{c}_{base} (black dotted line); backbone-binding positions, \bar{c}_{bb} (orange dotted line). The number of aligned sequences is given in brackets next to the family name.

clearly shows that scores for individual interacting residue positions vary considerably. In highly-specific and in non-specific families, interacting positions are heavily biased towards the top end of the distribution and there is usually a cluster of very conserved residue positions at 90–100. However, about a quarter of interacting residue positions score below the mean for the protein surface. In multi-specific families, interacting residue positions are biased towards the centre of the distribution. There is a broader spread of scores between 50 and 100, and a significant number of interacting positions score below 20 (i.e. these are variable).

Backbone-contacting residue positions

The variations in the conservation of individual interacting residue positions is best understood if we consider the interactions with DNA base and backbone groups separately. We start with backbone interactions, which we stated as acting mainly to stabilise the complex.⁴ As with all

interacting amino acid residues, backbone-contacting positions (\bar{c}_{bb}) tend to be better conserved than the rest of the protein surface (Table 3). In addition, highly specific and non-specific families have higher average scores than multi-specific families.

Individual backbone-contacting positions display quite a broad spread of scores in most families (Figures 2–4). In Figure 5(a), we plot the scores of individual backbone-contacting positions (c_{bb}) against the surface means for each family (\bar{c}_{surf}). As a crude summary, we consider positions scoring more than one standard deviation above the surface mean ($c_{int} > \bar{c}_{surf} + \sigma_{surf}$) as well conserved compared with the rest of the protein sequence. Those that score more than 1 SD below the surface mean ($c_{int} > \bar{c}_{surf} - \sigma_{surf}$) are considered poorly conserved. This provides a simple way to separate the well-conserved and the poorly conserved backbone-contacting positions, as there is no obvious cut-off between them (Figure 5(a)).

There are a total of 359 backbone-contacting positions. 251 (69.9%) score above the surface

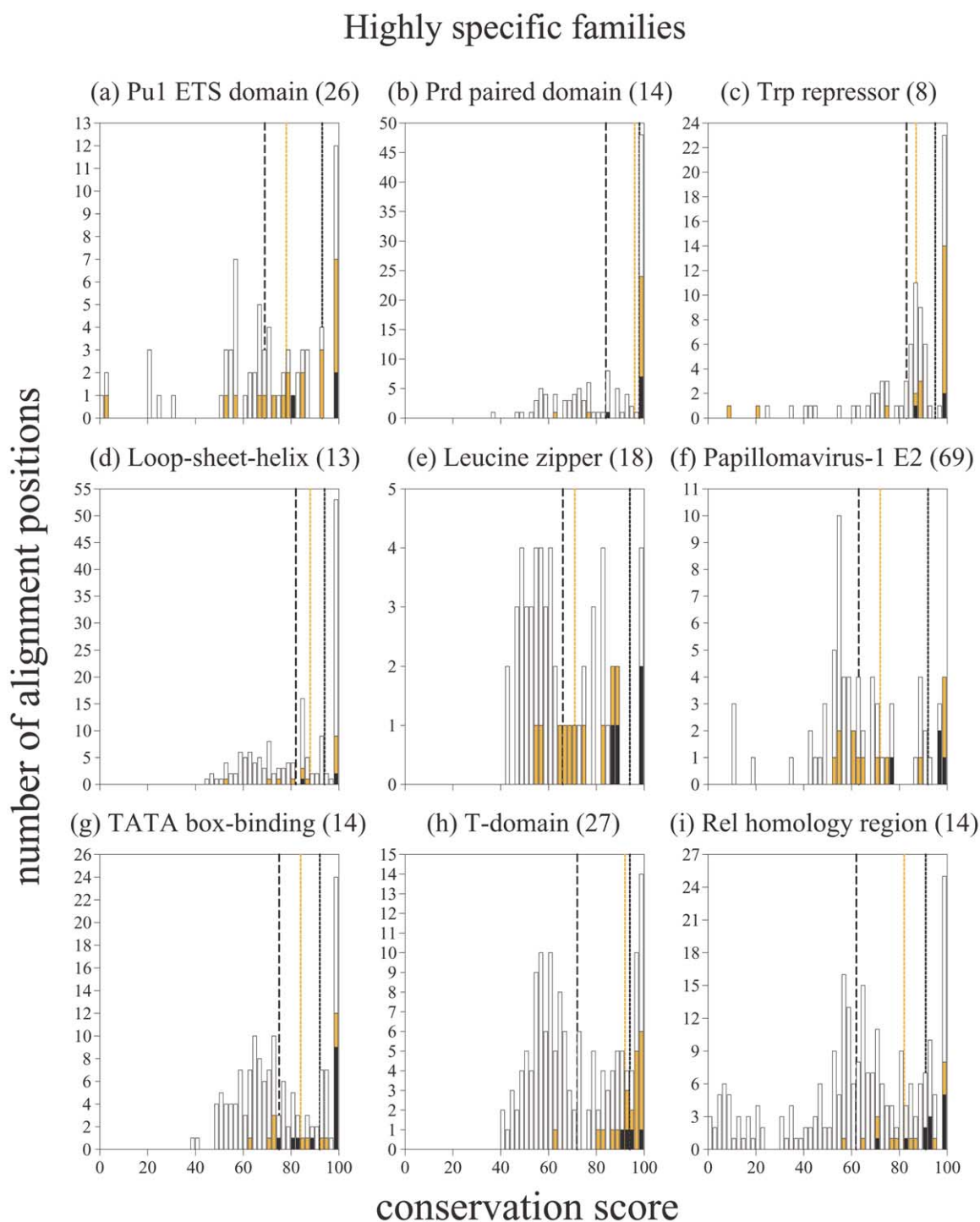


Figure 3. Conservation score distributions of surface alignment positions for highly specific families.

mean, and of these, 111 (30.9%) are well conserved. Only 22 positions (6.1%) are poorly conserved. That the former outnumber the latter by a factor of 5 indicates that backbone-contacting positions are generally very highly conserved.

Every protein family typically contains 5–15 well-conserved positions regardless of binding class (Figures 2–4). These positions provide a common framework of stabilising interactions for all member proteins in a family, and represent the minimum requirement of interactions for complex

formation. To test whether a high score actually corresponds to a conserved interaction, we examined 55 amino acid residue positions from families that contain multiple PDB structures. Of these, 45 (81%) positions interact with equivalent DNA backbone groups in over half the structures, so confirming that conserved amino acid residues also maintain similar interactions.

Many families also have one or two poorly conserved amino acid residues positions, although we emphasise that these are much less common than

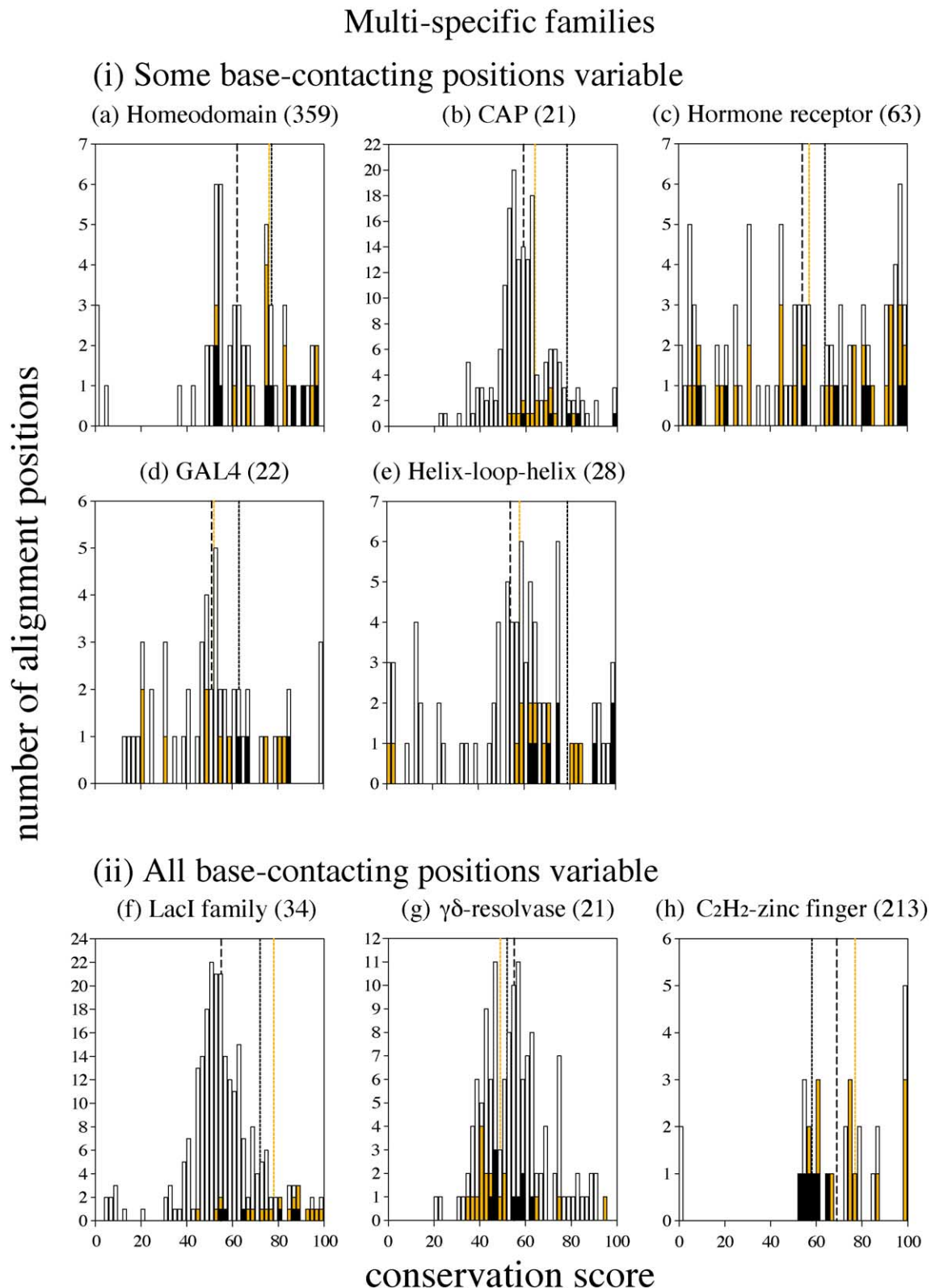


Figure 4. Conservation score distributions of surface alignment positions for multi-specific families.

high-scoring ones (Figures 2–4). These positions correspond to backbone interactions found only in a small subset of family members, or even just individual complexes. They are probably not essential for minimal DNA-binding because

most family members form complexes without them.

Significantly, the well-conserved and the poorly conserved amino acid residues are found on different regions of the protein surface. Well-conserved

Table 3. Mean conservation scores of alignment positions for each family

Family	Mean conservation scores for each family					
	All/ \bar{c}_{all}	Surface/ \bar{c}_{surf}	All inter-acting/ \bar{c}_{int}	Backbone/ \bar{c}_{bb}	Base/ \bar{c}_{base}	
Non-specific						
IHF	1ihfA	59.4	61.1 (21.1)	67.6	68.3	72.5
Polymerase Taq	1tauA	60.2	61.2 (16.4)	81.8	80.3	87.0
Polymerase- β	1bpyA	66.1	68.7 (16.2)	71.7	75.0	59.3
Dnase I	2dnjA	81.9	78.5 (13.8)	87.4	90.5	75.5
Mean of means/ \bar{C}		66.9	67.4	77.1	78.5	73.5
Highly specific						
Pu1 ETS domain	1pueE	67.8	69.6 (24.1)	79.9	78.9	93.5
Prd paired domain	1pdnC	85.2	84.4 (13.4)	97.0	96.9	98.2
Trp repressor	1trrA	84.4	83.0 (18.9)	88.4	87.3	95.4
Loop-sheet-helix	1tsrB	84.5	82.6 (17.2)	89.5	88.4	94.8
Leucine Zipper	2dgcA	65.9	66.1 (16.3)	78.2	71.9	94.0
Papillomavirus-1 E2	2bopA	64.3	63.3 (19.7)	73.1	72.9	92.8
TBP	1ytbA	76.4	75.2 (16.2)	89.5	84.2	94.3
T-domain	1xbrA	74.8	72.8 (17.7)	91.3	92.6	94.5
RHR	1nfkA	65.2	62.4 (26.5)	84.6	82.2	93.0
Mean of means/ \bar{C}		74.2	73.3	85.7	83.9	94.5
Multi-specific						
Homeodomain	1yrnA	63.8	62.1 (22.7)	76.2	76.4	77.8
LacI	1wetA	58.6	55.8 (16.3)	75.7	78.1	72.5
CAP	1berA	60.6	59.6 (13.0)	69.0	64.9	78.3
$\gamma\delta$ -resolvase	1gdtA	57.9	55.7 (14.0)	50.6	49.6	52.6
C ₂ H ₂ -zinc finger	1aayA1	69.6	69.4 (22.6)	71.5	77.8	58.0
Hormone receptor	2nllA	56.0	54.7 (29.5)	59.3	57.7	64.4
GAL4	1d66A	51.3	51.4 (22.2)	53.5	52.5	63.3
HLH	1hloA	54.5	54.5 (25.7)	66.8	58.0	79.9
Mean of means/ \bar{C}		59.0	58.8	65.3	64.4	68.4
Overall mean/ \bar{C}		67.0	66.7	76.3	75.4	80.3

The first set of columns gives mean scores of the whole alignment (\bar{c}_{all}), positions on the protein surface (\bar{c}_{surf}), and positions in contact with the DNA (\bar{c}_{int}). The second set divides interacting positions into those that contact DNA bases (\bar{c}_{base}), and those that contact the backbone only (\bar{c}_{bb}). A mean of means (\bar{C}) is calculated for each binding class, and the overall mean (\bar{C}) is calculated for all 21 families.

residue positions usually reside in DNA-binding motifs. These are small, well-defined structural units found in many families that are used for much of the DNA-binding, for example the helix-turn-helix and C₂H₂-zinc finger motifs.^{11,12} Of the 111 conserved positions, 67 are in a helix-turn-helix motif, zinc-coordinating motif, major groove-binding α -helix or DNA-binding β -strands. The remaining 44 positions reside in loop regions, most of which are from proteins that employ complex loop networks for binding such as the RHR and enzymatic families.¹¹

Poorly conserved residue positions are located mostly in loop regions outside defined DNA-binding motifs: (i) N or C-terminal ends of the alignment; or (ii) inter-domain linkers. The first are

found mostly in families whose sequence alignments are incomplete at the N and C-terminal ends, e.g. Trp repressor. Although we cull the alignments to minimise such effects, there are inevitably some short protein sequences that distort the conservation scores in these regions. The hormone receptor family is an interesting exception that actually uses different-length protein sequences to alter binding characteristics. Five interacting positions reside in a functionally important C-terminal α -helix (called A-helix) that is present in only a subset of members (e.g. thyroid receptors). The additional amino acid residue-backbone contacts allow receptors containing the A-helix to function as monomers as well as dimers.^{26–28} The second set of poorly conserved

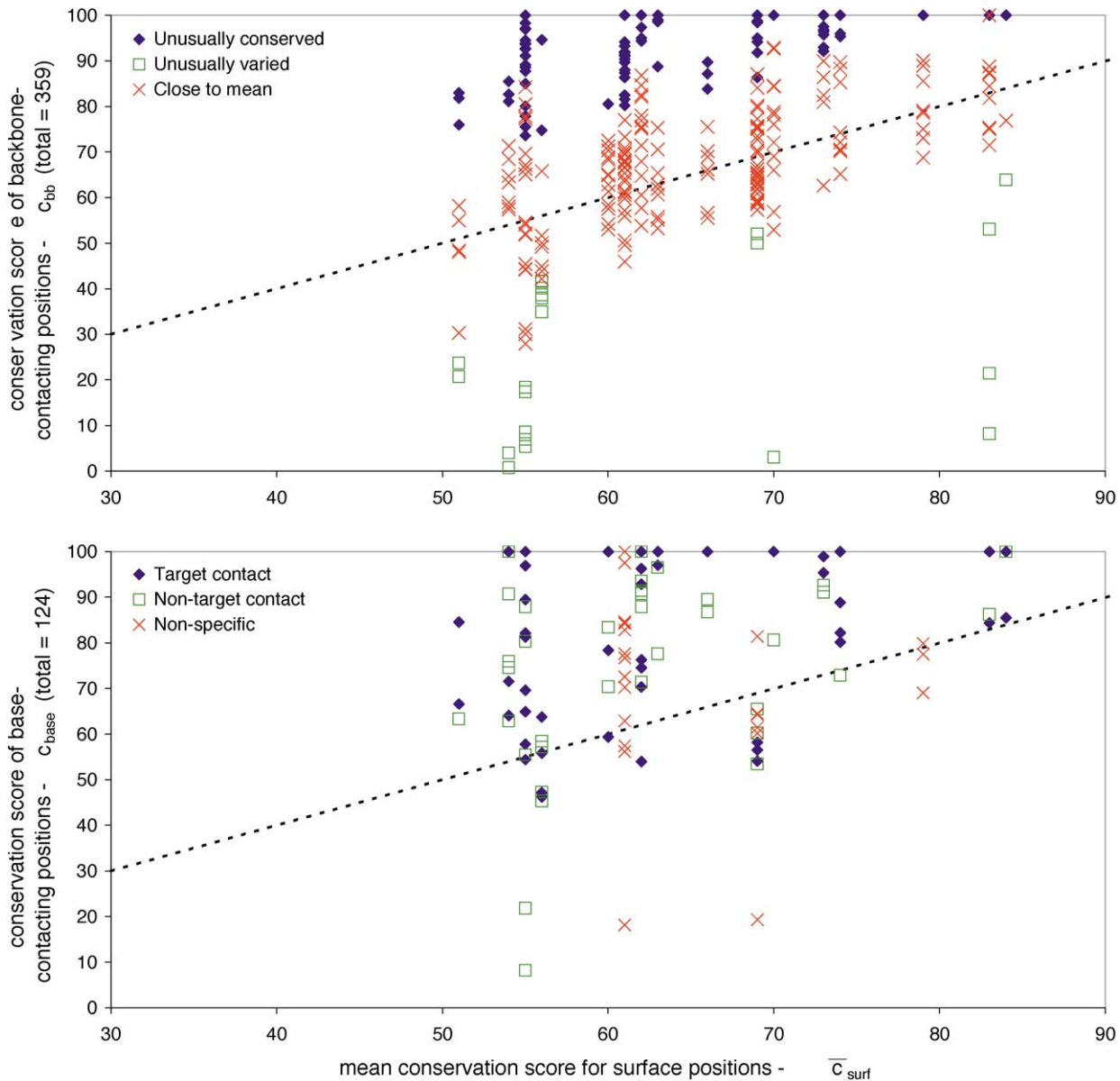


Figure 5. Scatter plots of conservation scores of individual interacting positions (c_{int}) against the mean score for all surface positions (\bar{c}_{surf}) for each family. (a) Plot for backbone-contacting positions. Points are coloured according to level of conservation: (\blacklozenge) >1 SD above the mean, (\square) >1 SD below the mean, (\times) within 1 SD of the mean. (b) Plot for base-contacting positions. Points are coloured according to whether they contact (\blacklozenge) a target base or (\square) a non-target base, or (\times) if from a non-specific family.

residues reside in loop regions that act as linkers between structural domains that contact different regions of the DNA. They make happenstance contacts with the DNA backbone while spanning the domains, e.g. the Rap1 family.^{29,30}

In summary, all protein families, regardless of their binding class, contain a core of well-conserved backbone-contacting residue positions that provide a common framework of stabilising interactions. These positions are normally found within the well-defined DNA-binding motif that characterises the family. Most families also contain a few poorly conserved positions outside the motifs, and these provide happenstance interactions in individual protein–DNA complexes.

Base-contacting residue positions

Next we turn to base interactions, which, in addition to stability, we consider to provide most of the target sequence recognition.⁴ Again, base-contacting residue positions (\bar{c}_{base}) are generally better conserved than the rest of the protein surface (Table 3). However, in significant contrast to interactions with the DNA backbone, there is a large difference in the conservation of base-contacting positions between binding classes.

Base-contacting positions are extremely well conserved for highly specific families ($\bar{c}_{base} = 91.3 - 98.2$), where similar amino acid residues are retained at base-contacting positions (Figure

Table 4. Sets that group together amino acids according to their binding preferences with bases¹¹

Set	Amino acids	Mode of interaction	Recognised base
Hydrogen bond			
(a)	[ARG, LYS]	Multiple-donor	G/complex
(b)	[HIS]	Multiple-donor (bifurcate)	G
(c)	[SER]	Multiple-donor (bifurcate)	G
		Acceptor+donor	complex
(d)	[ASN, GLN]	Acceptor+donor	A/complex
(e)	[ASP, GLU]	Multiple-acceptor	complex
van der Waals contacts			
(f)	[PHE, PRO]	Ring-stacking	A, T
(g)	[THR]	Methyl contact	T
(h)	[GLY, ALA, VAL, LEU, ISO, TYR]	-	many (non-specific)
		-	-
No base contact			
(i)	[CYS, MET, TRP]	-	-

Sets (a)–(e) comprise amino acid residues that bind predominantly through hydrogen bonds, sets (f)–(h) are those that bind bases *via* van der Waals contacts, and set (i) includes those that do not interact with bases in significant numbers. Each set has a brief summary of the nature of interaction that can be achieved using the side-chains, and the bases that are preferred for binding.

3). This allows member proteins to target the same DNA sequences. The conservation score distributions in Figure 3 show that almost all base-contacting positions score above 80. The only exceptions are found in the TATA box-binding and RHR families, which contain six or seven residue positions that score lower (see below).

Residue positions are much less conserved for multi-specific families ($\bar{c}_{\text{base}} = 52.6 - 78.3$). By mutating amino acid residues at base-contacting positions, different members of a family can recognise distinct target sequences while using homologous DNA-binding structures. We can divide the conservation score distributions into two types (Figure 4). (i) First are families that have two or three conserved amino acid residue positions ($c_{\text{base}} > 80$), and a similar number of less conserved positions ($c_{\text{base}} = 50 - 80$). We anticipate that these families use the conserved protein residues to recognise a “core” segment of the target sequence that is common for all family members, and the mutated protein residues to differentiate between segments of the target sequences that vary (homeo-domain, CAP, hormone receptor, GAL4 and helix-loop-helix; Figure 4(a)–(e)). (ii) Second are families whose base-contacting positions are all variable ($c_{\text{base}} = 40 - 60$). We expect that members of these families target completely different DNA sequences (LacI, $\gamma\delta$ -resolvase, C₂H₂-zinc finger; Figure 4(f)–(h)).

Finally, although non-specific proteins bind independently of DNA sequence, families nonetheless contain several base-contacting residues ($\bar{c}_{\text{base}} = 52.8 - 87.0$). These have variable levels of conservation (Figure 2), and we find that the con-

served residue positions provide crucial interactions for the function and stability of these protein complexes.

We will now discuss different aspects of these base-contacting positions in greater detail. Further descriptions of individual families are available in Appendix 1.

Universal amino acid-base interactions

Before plunging into the details of base interactions, it is useful to describe briefly the concept of “universal” amino acid-base specificity. Previously, we and others analysed the interactions in protein–DNA complex structures in the PDB to examine amino acid-base-pairs that are favoured across all protein–DNA families.^{2–4} The most prominent interactions taken from 129 complexes are listed in Table 4. Generally, hydrogen bonds display the strongest specificity: these include the interactions of arginine, lysine, histidine and serine with guanine, asparagine and glutamine with adenine. In addition, amino acid residues with more than one side-chain hydrogen-bonding atoms can produce “complex” interactions with multiple bases simultaneously; for example arginine and lysine with two adjacent bases presenting acceptor atoms. Though weaker, we found favourable amino acid-base-pairings using van der Waals contacts; in particular, the interactions of proline and phenylalanine residues with thymine or adenine. Clearly these pairings do not exclude other types of interactions from providing specificity in particular circumstances (context-dependent specificity). However, in our current study,

we find that a large proportion of specific interactions involve these universal pairings, and the effect of amino acid mutations are best understood in terms of these pairs.

Special case: non-specific families

We first examine the amino acid-base contacts in non-specific families. All families have several highly conserved base-contacting positions, despite the absence of sequence recognition (Figure 2). However, none follow the universal pairings, and interactions are always made *via* the minor groove. Here, bases have relatively similar van der Waals surfaces and present similar hydrogen bond acceptors (O2 on purine and N3 on pyrimidine). Thus, interactions can be made non-specifically to the four base types using either hydrogen bonds or van der Waals contacts.³¹

Regardless, the conserved base-contacting residues perform two important functions. (i) First, they stabilise the deformed DNA structure, which often has a widened minor groove. For example, in the IHF family, interactions are made using a β -hairpin DNA-binding motif. An absolutely conserved proline residue (PDB code 1ihf, chain A residue 65) intercalates between base-steps, producing a sharp kink in the DNA.²⁵ Further positions in the flanking β -strands help maintain the local distortion of the DNA: arginine residues bind minor groove acceptors (1ihf, chain A residues 60 and 63) and hydrophobic residues produce van der Waals contacts (1ihf, chain A residues 62, 64 and 73). Similar interactions are found in the DNase I family, where conserved polar residues widen the minor groove (see Appendix 1).²⁴ (ii) Second, base contacts provide important interactions in the enzyme active site of polymerase families. In the Taq polymerase, two polar residues (1tau, chain A residues 746 and 754) bind the O2 and N3 acceptor atoms from a purine-pyrimidine base-pair. The base atoms occupy identical locations in all four correct combinations of Watson-Crick base-pairs, but not for non-Watson-Crick pairs. Thus, the interactions provide a sequence-independent method for checking that correct nucleotide pairs are being incorporated into the DNA at the polymerase active site.³² In addition, aromatic amino acid residues (1tau, chain A residues 667 and 671) stabilise newly added bases through ring-stacking interactions, and polar residues (1tau, chain A residues 583) maintain the widened conformation of the DNA minor groove. Similar interactions are found in the polymerase- β family (see Appendix 1).

Sequence-specific families

Having discussed the base contacts in non-specific families, we now shift our focus to the families that bind DNA specifically.

Types of amino acid-base interactions

Target sites are defined as the consensus DNA sequence to which proteins bind specifically (see Materials and Methods). With all sequence-specific families, we can separate interactions into those with bases inside the target sequence and those with bases outside the target sequence. A total of 69 base-contacting residue positions interact inside the target sequence and 35 interact outside. This corresponds to two to five target-contacting and one to five non-target-contacting positions per family. The fact that there are twice as many target-contacting residue positions emphasises the role of amino acid-base interactions for target recognition.

Interactions within the target are probably the most important for sequence recognition. These tend to involve direct hydrogen bonds between the amino acid and base, though there are notable exceptions, such as the TBP family, where aromatic protein residues play a central role in specificity (see Appendix 1). In both highly specific and multi-specific families, the most common pairings are arginine/lysine-guanine (20 positions) and asparagine/glutamine-adenine (four positions). We observe a large number of complex hydrogen bonds with multiple base-steps (19 positions).

Non-target-contacting residue positions generally produce van der Waals or water-mediated contacts. These are effectively equivalent to backbone contacts, providing stability to the complex, but little specificity. Most of the protein residues are hydrophobic (25 positions), but several positions contain polar or charged amino acid residues that fail to produce hydrogen bonds with bases (ten positions); these interact predominantly with the DNA backbone, and the interactions with bases are secondary to their stabilising role.

We plot the conservation scores of individual base-contacting positions (c_{base}) against the surface means for each family (\bar{c}_{surf}) in Figure 5(b). On average, target-contacting positions ($\bar{c}_{\text{base}} = 82.5$) are better conserved than non-target-contacting positions ($\bar{c}_{\text{base}} = 74.4$). But the plot clearly shows that data points for target and non-target-contacting positions are interspersed and thus hard to differentiate only by conservation.

In highly specific families, 26 out of 38 target-contacting positions are conserved absolutely. At the remaining positions, mutations are not expected to alter specificity because they involve amino acid residues with similar binding preferences; for example, arginine to lysine (see Appendix 1). In multi-specific families, 13 of the 31 target-contacting positions are well conserved ($\bar{c}_{\text{base}} > \bar{c}_{\text{surf}} + \sigma_{\text{surf}}$), which we expect to provide recognition for conserved regions of the target sequence. In all, 18 residue positions are variable ($\bar{c}_{\text{base}} < \bar{c}_{\text{surf}} + \sigma_{\text{surf}}$), translating to two or three target-contacting residue positions in each family. The effects of amino acid mutations at these residue positions are described below.

Fuzzy recognition of target sequences

There are two levels at which DNA-binding proteins recognise different DNA sequences. The first is through fuzzy recognition of different target sites by a single protein sequence. The second is through amino acid mutations at target-contacting positions.

Fuzzy recognition is best observed in highly specific families. Despite conservation of base-contacting positions, there are often variable base-pairs within the consensus binding sequence. Target sequences are typically four to eight base-pairs long, of which two to four vary. This means that individual proteins can usually bind a range of different targets with similar affinity (see [Figure 1](#)). For example, leucine zippers bind to sequences -TGnn- and -CAnn-, while TBP's recognise -TATA(T/A)A(T/A)- (see [Appendix 1](#)).

There are two contributing factors. First, proteins generally contact two to five bases in the target sequence, so they do not interact with all the base-pairs. Base-pairs are almost always variable if there is no contact with the protein ([Appendix 2](#)) and recent work has shown that the conservation of base-pairs in a consensus binding sequence is highly dependent on the number of interactions they make with the protein.³³

The second factor is more surprising, as variations occur at base-pairs in contact with the protein; these are explained by the ability of protein and DNA residues to interact in different ways. From the 12 examples in our dataset, we observe three patterns of interactions. (i) Bases offer equivalent interacting atoms, so making them indistinguishable. In the leucine zipper family, the asparagine side-chain (2dgc, chain A residue 235) can rotate to recognise both -TC'- and -CT'- (cytosine and thymine are positioned diagonally on opposite strands). (ii) Amino acid residues can form distinct bonds depending on the base that is presented. In the LSH family, cysteine (1tsr, chain A residue 277) acts as a donor or acceptor depending on whether thymine or cytosine is present. (iii) The protein and DNA accommodate their three-dimensional structures to maintain similar interactions. In the Prd paired domain, a peptide O atom (1pdn, C15) interacts with guanine regardless of the strand as long as it is in the equivalent base-step. The protein residue is in a loop region that adapts its shape according to the base sequence.

Amino acid mutations and differential sequence recognition

Fuzzy recognition allows a single protein to bind variable consensus DNA sequences. However, they do not allow multiple proteins to differentiate between distinct target sequences. Multi-specific families achieve this by mutating amino acids at target-contacting positions. Details of mutations in each protein family are provided in [Appendix 1](#).

Of the 18 variable target-contacting positions, almost all involve mutations of hydrogen-bonding amino acid residues. These mutations are likely to alter the target specificity most. The most common substitutions are arginine, lysine, and serine (seven examples), which interact mainly with guanine, to asparagine or glutamine that favour adenine, or to hydrophobic residues that have little binding preference. Significantly, none of the mutations involve interactions that use complex hydrogen bonds to multiple bases. The relative complexity of these interactions probably makes them less amenable to change by amino acid mutations.

The effects of mutations are less predictable when interactions are non-universal (five positions). For example, in the hormone receptor family, glutamate-cytosine specificity is lost on introduction of glycine, which lacks hydrogen-bonding capability (2nll, chain B residue 321). At another position, substitution of glycine or alanine with a larger valine residue introduces specificity for thymine (2nll, chain B residue 325). Of course, these changes can be understood when individual structures are inspected visually. However, their specificities are context-dependent, and the outcomes of mutations cannot be predicted readily for all protein-DNA complexes. Further difficulty is encountered when aligned protein residues contact different base-pair positions in the target sequence, as observed in the C₂H₂-zinc finger and homeodomain families. Therefore, although broad conclusions can be made about the effects of amino acid mutations, it is difficult to predict the precise effect on specificity without careful study of individual families.

Recognition subfamilies and regulatory functions

Given the presence of mutations that lead to differential DNA sequence recognition, each multi-specific family can be divided into subfamilies. Members of the same subfamily have identical or equivalent amino acid residues (i.e. protein residues with similar recognition properties, [Table 4](#)) at interacting positions, and are expected to target the same DNA sequences. Members of different subfamilies vary at one or more target-contacting positions and are expected to recognise distinct DNA sequences.

Although subfamily classifications are based on the amino acid sequences at target-contacting positions, they correspond well with differential specificities for distinct DNA sequences. By inspecting families containing more than one PDB structure (homeodomain, C₂H₂-zinc finger, hormone receptor, HLH families), we found that proteins always belong to the same subfamily if complexed with the same DNA sequences, and to different subfamilies if complexed with different sequences (see [Appendix 2](#)).

There are between three and 42 subfamilies in each alignment. Thus, protein families can be

separated into those that contain many subfamilies (homeodomain, LacI, C₂H₂-zinc finger, and $\gamma\delta$ -resolvase each have >20 subfamilies) and those that are less variable (CAP, hormone receptor, GAL4 and helix-loop-helix each have less than four subfamilies). This underlines our observation that certain DNA-binding motifs are more versatile and can accommodate amino acid mutations more readily.

The significance of subfamilies extends beyond the recognition of distinct DNA sequences, and we observe that they often correspond to the different regulatory functions that have evolved for each family of transcription factors. Generally, proteins of identical regulatory functions (e.g. orthologues) belong to the same subfamily, while those of differing functions (e.g. paralogues) belong to separate ones.

This is best illustrated by the example of the CAP family (Table 5). Equivalent Tables for the remaining families are included in Appendix 1. The CAP family comprises proteins that use the helix-turn-helix motif for recognition of the target DNA sequence. Two broad categories of proteins are included in the alignment: the CAPs, which control transcription of catabolite-sensitive operons, and the fumarate/nitrate regulators, which control genes linked to anaerobic electron transport systems. The structure of only CAP is available in the PDB.

Table 5(a) displays a multiple alignment of the family, with residue positions highlighted according to their interactions with the DNA backbone and bases. We plot the conservation score for each position along the bottom of the alignment. Table 5(b) focuses on the positions that interact with the bases and divides the CAP family into subfamilies. Three positions (1ber, chain A residues 180, 181, and 185) interact with the target site, but indirect recognition of the sequence is important, as the bound DNA is deformed significantly in the complex.^{34–36} In CAP, arginine residues at positions 180 and 185 recognise guanine, and the glutamate residue at position 181 binds cytosine. Mutagenesis studies have shown that point mutations at these positions alter the specificity of the protein.^{37–39}

There are three subfamilies, which are defined by amino acid mutations at positions 180 and 181. All CAPs belong to subfamily 1, while the anaerobic regulators belong to subfamilies 2 and 3. Mutagenesis studies of *Escherichia coli* promoters have succeeded in converting a CAP-binding site to a fumarate/nitrate regulator site with a single substitution at the base-pair contacted by position 180.⁴⁰

Discussion

Here, we have combined data for over 150 PDB structures, 1500 SWISS-PROT sequences and target site information to investigate the molecular basis for DNA-binding specificity and how related proteins have evolved to recognise different target sites. The 21 protein families containing between five and 359 members were classified into one of three binding classes, each having a distinct pattern of amino acid conservation.

Caveats

Before summarising the main conclusions from the work, we must first assess the scope of our analysis with several important factors in mind.

First, although much of the specificity is explained by direct and water-mediated amino acid-base interactions, many DNA-binding proteins use indirect methods of sequence recognition. For example, the TBP and CAP families use DNA-bending extensively. Several studies have assessed the importance of such factors,¹³ however, it is difficult to include them consistently in a global analysis of diverse protein families, as they are still poorly understood except in individual protein–DNA complexes. Thus, we excluded such considerations from the current study.

Second, we assume that all family members use the same alignment positions to interact. Several members of the C₂H₂-zinc finger and homeodomain families (see Appendix 1) show that aligned positions do not always interact with equivalent bases in the target sequence. However, in comparing the target-contacting positions to those that interact in >50% of structures, we observe that many of the interactions are actually made with equivalent amino acid residues. Therefore, as this problem is found only in two families, which are also the largest, a possible solution in future studies may be to separate them into smaller families with members that are more similar to each other.

Third, we focus on sequence recognition by single DNA-binding subunits. For single-domain proteins, this corresponds to the whole target sequence. However, further studies should consider implications of proteins with multiple binding subunits: the combinations of half-site sequences for heterodimers (e.g. the leucine zipper family), their relative orientations (hormone receptors bind direct or inverted repeats depending on dimerisation) and the spacing between them (GAL4 dimers differentiate the spacings between

In (a) the alignment is culled between residues 60 and 135 as there are no interactions with the DNA in this region. Residue positions are coloured according to whether they interact with the DNA backbone (orange) or bases (black). The conservation score for each position is plotted along the bottom of the alignment. In (b) alignment positions are numbered according to the residue positions in the representative structure. Positions that interact with the target sequence are highlighted in bold. For each protein, we provide the one-letter code (upper-case) and amino acid set (lower-case) of the residue found at the alignment positions. Member proteins are divided into subfamilies according to the amino acid sequence at the target-contacting positions.

subsites). Additional variability may be introduced by changing any of these parameters.

Finally, the assignment of binding classes is dependent largely on the members contained within the families, and the information about them that is available. Non-specific families are likely to remain so regardless of size, because sequence-independent binding is often a requirement for their functions. Obviously, multi-specific alignments will remain so unless members are split into smaller families. However, some of the highly specific families may become multi-specific as databases expand and new targets or new members with alternative specificities are found.

Summary of findings

Families can be classified using a three-class model for DNA-binding

The classes are: (i) non-specific, where binding is not dependent on the base sequence; (ii) highly specific, where binding is specific and all members of a family preferentially target the same or similar base sequences; and (iii) multi-specific, in which binding is also specific, but different members of a family target distinct base sequences. Four families were defined as non-specific, nine as highly specific and eight as multi-specific.

Amino acids that interact with the DNA are better conserved than those that do not

However, conservation scores for individual alignment positions vary significantly, and are often interspersed with non-interacting residues. Therefore, it is extremely unlikely that we could predict DNA-binding sites solely on the basis of amino acid conservation.

Sequence-specific families place greater emphasis on interactions with DNA bases than non-specific families

Between 14 to 47 residue positions contact the DNA in each family. The ratios of backbone to base-contacting positions are high for non-specific families and lower for the highly specific and multi-specific families. This indicates that sequence-specific families place greater emphasis on interactions with bases.

DNA backbone-contacting positions are well conserved in all families

These provide a common framework of interactions that stabilises the complex. Families typically contain 5–15 well-conserved backbone-contacting positions that provide happenstance interactions in individual complexes.

Conservation of base-contacting positions depends on the binding class of the family

Non-specific families contain several conserved base-contacting positions. Interactions are invariably in the DNA minor groove, where there is little discrimination between base types. In highly specific families, target-contacting positions are absolutely or highly conserved, although mutations between amino acid residues with similar binding preferences are often tolerated. In multi-specific families, base-contacting positions are frequently mutated.

Fuzzy recognition allows single proteins to recognise different, but related target DNA sequences

Proteins often tolerate changes in the target sequence without greatly affecting the binding specific, e.g. leucine zippers bind –TGnn- and CAnn-. This is because proteins rarely contact all base-pairs in the target sequence (i.e. non-interacting base-steps frequently vary), and there is some flexibility in the interactions between amino acid residues and bases (i.e. equivalent atomic interactions can be satisfied by different amino acid-base-pairs).

Members of multi-specific families recognise different DNA sequences by mutating amino acids at base-contacting positions

The mutations frequently involve amino acids that interact with favoured bases, and the effects of substitutions are relatively easy to predict. Each multi-specific family can be divided into subfamilies according to the amino acid residue sequences at base-contacting positions, where members of the same subfamily are predicted to bind the same DNA sequences and those of different subfamilies to bind distinct sequences. Rather than design a new DNA-binding structure to recognise a new DNA sequence, mutations of base-contacting positions offers an economic way of producing proteins that target different binding sites, and so regulate the expression of different genes. We found that some families, particularly the C₂H₂-zinc fingers and homeodomains, are much more versatile in their ability to accommodate mutations, and these families have duplicated prolifically to provide a large number of homologues in eukaryotic genomes.

Conclusion

In conclusion, we observed that combined analysis of sequence and structural data provides invaluable insight into the method of target sequence recognition by DNA-binding proteins. We emphasise that, while there are many principles that apply generally, each family must be

studied in great detail in order to understand the complexities of how they recognise their target sequence. Of particular interest is how proteins with similar DNA-binding scaffolds achieve different specificities by altering their amino acid residue sequences. In doing so, proteins have evolved distinct functions, therefore allowing enzymes to act at different sites and transcription factors to regulate expression of different genes. It appears that some structural motifs, especially the very simple, but stable ones, are better designed to accommodate these mutations than others.

The datasets we have produced provide a large collection of protein sequences and their corresponding binding sites, were known. Much of the data that we have produced are available from the web as Supplementary Material. They may be of use in further research to predict the DNA sequence of recognition sites for novel protein sequences within each family.

Materials and Methods

The following procedure was used to construct the multiple sequence alignments and calculate the amino acid conservations of the 21 protein families. (i) Structural families of protein–DNA complexes were compiled from the January 2000 release of the PDB.^{21,22} (ii) The structural families were expanded to include sequence homologues from SWISS-PROT.²³ (iii) Multiple sequence alignments of the expanded families were produced. (iv) Conservation scores were calculated for each alignment position. (v) Finally, the surface and interacting alignment positions were identified in the protein structures, and extrapolated across all family members. Each step is described in more detail below.

Protein structural families from the Protein Data Bank

Protein–DNA complex structures solved by X-ray crystallography to a resolution greater than 3.0 Å were obtained from the January 2000 release of the PDB. 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 in length. From this set, we excluded structures containing single-stranded and quadruple-stranded DNA. This resulted in a dataset of 240 protein–DNA complexes.

The PDB entries were classified into structural families by comparing their structures in pairs using the structure alignment program SSAP.⁴¹ This uses a dynamic programming method⁴² and assesses the similarity between proteins by comparing the structural environments of the constituent amino acid residues. SSAP returns a score of 100 for identical proteins, and >80 for homologues; proteins are assigned automatically to the same family if they score above this cut-off. More distantly related proteins that give scores of >70 are placed in the same family if they perform similar biological functions.⁴³ The result is a total of 54 protein families and multiple structural alignments of each family were made using the CORA program suite.⁴⁴

Proteins were broken down into their constituent DNA-binding domains before conducting the align-

ments. 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 C₂H₂-zinc fingers, however, a chain contains several binding domains; in such cases, the subunits were separated into the appropriate segments. Therefore, in the multiple alignments we describe below, each sequence contains just one DNA-binding domain. In Table 1, we list the number of structures contained within the 21 protein families under consideration in this study. A complete list of all PDB entries, structural classifications and a description of each homologous family is available from a previous publication.¹¹

Extracting sequence homologues from SWISS-PROT

The HMMER program suite[†] was used to construct hidden Markov models (HMM) for the structural alignments of each family and then to search for homologues in SWISS-PROT (release 39.0). Default program settings permitted matches that were global with respect to the HMMs and sequences, i.e. the program found only complete protein segments defined by the models and only single hits were allowed within the sequences searched. The outputs were inspected manually to remove proteins with non-DNA-binding or hypothetical functions based on the SWISS-PROT annotations. Matches with >95% sequence identity were discarded and a single representative, the protein with greatest similarity to the discarded members of the group, was retained for the non-redundant dataset. Families with less than five members were excluded from the study, as they comprise near-identical proteins and do not display sufficient sequence divergence. The number of SWISS-PROT sequences and final non-redundant total for each family is given in Table 1.

Building multiple sequence alignments

Multiple alignments were built for each sequence family using the original HMMs as templates. As proteins are variable in length, alignments were cleaved at the N- and C-terminal ends to remove sequence segments that extend beyond the regions defined by the models.

Amino acid conservation

The PET91 substitution matrix (Appendix 2) was used to score amino acid conservation in each alignment.⁴⁵ Each cell in the matrix represents a pairwise score, normalised between 0 (unconserved) and 100 (conserved), for the substitution of one amino acid type by another. Amino acid conservation at each multiple alignment position was calculated as the unweighted average score of all pairwise comparisons for that position. Gaps in the alignment scored 0.

Surface and interacting alignment positions

For each family, we identify alignment positions that lie on the protein surface and those that interact with the DNA. Only alignment positions on the protein surface are considered, because we are interested in amino acid residues that potentially can interact with DNA.

[†] <http://hmmer.wustl.edu>

These are defined as those exposing >5% of the amino acid residue in at least one aligned structure. Solvent-accessibilities are calculated for every structure using the program NACCESS† after removing the DNA from the PDB files. Positions that interact with the DNA are defined as those that contact the DNA in one or more complex, regardless of whether the structure is included in the non-redundant dataset. Hydrogen bonds, van der Waals contacts and water-mediated bonds are considered, and the method for calculating protein–DNA described by Luscombe *et al.*⁴ Assuming direct read-out, positions are divided into those that contact the DNA backbone only, and those that contact bases or both base and backbone groups.

Target sequences and binding site classifications

The target DNA sequences were obtained from the literature. Where possible, we distinguish between the cognate (biological) and non-biological binding sites of proteins. A list of the target sequences for all the families is available in the Supplementary Material. While it is not possible to obtain target site definitions for all protein sequences, we have defined the binding class of each family on the basis of the available data.

Binding-class assignments are unambiguous for the four non-specific families as the sequence-independent nature of binding is integral to their protein functions. Although the IHF and DNase I families do display slight sequence-dependence through indirect read-out of the local DNA structure, they are considered non-specific because their preferences are not strong enough to define clear targets.^{24,25} Highly specific families are defined because only single, or very similar target sites are found for most of the member proteins. Multi-specific families are defined because different members are associated with two or more distinct target sequences.

Supplementary material

Supplementary Material pertaining to this paper is available from <http://bioinfo.mbb.yale.edu/~nick/proteinDNA>.

Acknowledgments

We thank Roman A. Laskowski for useful comments on the manuscript. N.M.L. was supported by the BBSRC Special Studentship and is currently sponsored by the Anna Fuller Fund. This is a publication of the Bloomsbury Centre for Structural Biology.

References

- Pabo, C. O. & Necludova, L. (2000). Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597–624.
- Suzuki, M. (1994). A framework for the DNA–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317–326.
- Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.
- Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860–2874.
- Suzuki, M. & Yagi, N. (1994). DNA recognition code of transcription factors in the helix–turn–helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.
- Suzuki, M., Gerstein, M. & Yagi, N. (1994). Stereochemical basis of DNA recognition by Zn fingers. *Nucl. Acids Res.* **22**, 3397–3405.
- Suzuki, M. & Gerstein, M. (1995). Binding geometry of α -helices that recognize DNA. *Proteins: Struct. Funct. Genet.* **23**, 525–535.
- Suzuki, M., Yagi, N. & Gerstein, M. (1995). DNA recognition and superstructure formation by helix–turn–helix proteins. *Protein Eng.* **8**, 329–338.
- Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877–896.
- Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
- Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. (2000). An overview of the structures of protein–DNA complexes. *Genome Biol.* **1**, REVIEWS001, 1–10.
- Harrison, S. C. (1991). A structural taxonomy of DNA-binding domains. *Nature*, **353**, 715–719.
- Dickerson, R. E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucl. Acids Res.* **26**, 1906–1926.
- Choo, Y. & Klug, A. (1997). Physical basis of a protein–DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117–125.
- Grishin, N. V. & Phillips, M. A. (1994). The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci.* **3**, 2455–2458.
- Valdar, W. S. & Thornton, J. M. (2001). Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Struct. Funct. Genet.* **42**, 108–124.
- Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
- Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). Evolutionarily conserved $\alpha\beta\gamma$ binding surfaces support a model of the G protein–receptor complex. *Proc. Natl Acad. Sci. USA*, **93**, 7507–7511.
- Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27.
- Lichtarge, O., Yamamoto, K. R. & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325–337.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R. *et al.* (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**, 319–324.

† <http://wolf.bms.umist.ac.uk/naccess/>

22. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
23. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
24. Lahm, A. & Suck, D. (1991). DNase I-induced DNA conformation: 2 Å structure of a DNase I-octamer complex. *J. Mol. Biol.* **222**, 645–667.
25. Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. (1996). Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell*, **87**, 1295–1306.
26. Wilson, T. E., Fahrner, T. J., Johnston, M. & Milbrandt, J. (1991). Identification of the DNA binding site for NGFI-B by genetic selection in yeast. *Science*, **252**, 1296–1300.
27. Wilson, T. E., Paulsen, R. E., Padgett, K. A. & Milbrandt, J. (1992). Participation of non-zinc finger residues in DNA binding by two nuclear orphan receptors. *Science*, **256**, 107–110.
28. Rastinejad, F., Perlmann, T., Evans, R. M. & Sigler, P. B. (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature*, **375**, 203–211.
29. König, P., Giraldo, R., Chapman, L. & Rhodes, D. (1996). The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell*, **85**, 125–136.
30. Schwabe, J. W. & Rhodes, D. (1997). Linkers made to measure. *Nature Struct. Biol.* **4**, 680–683.
31. Moravek, Z., Neidle, S. & Schneider, B. (2002). Protein and drug interactions in the minor groove of DNA. *Nucl. Acids Res.* **30**, 1182–1191.
32. Georgiadis, M. M., Jessen, S. M., Ogata, C. M., Telesnitsky, A., Goff, S. P. & Hendrickson, W. A. (1995). Mechanistic implications from the structure of a catalytic fragment of Moloney murine leukemia virus reverse transcriptase. *Structure*, **3**, 879–892.
33. Mirny, L. A. & Gelfand, M. S. (2002). Structural analysis of conserved base pairs in protein–DNA complexes. *Nucl. Acids Res.* **30**, 1704–1711.
34. Parkinson, G., Wilson, C., Gunasekera, A., Ebright, Y. W., Ebright, R. E. & Berman, H. M. (1996). Structure of the CAP–DNA complex at 2.5 Å resolution: a complete picture of the protein–DNA interface. *J. Mol. Biol.* **260**, 395–408.
35. Chen, S., Vojtechovsky, J., Parkinson, G. N., Ebright, R. H. & Berman, H. M. (2001). Indirect readout of DNA sequence at the primary-kink site in the CAP–DNA complex: DNA binding specificity based on energetics of DNA kinking. *J. Mol. Biol.* **314**, 63–74.
36. Chen, S., Gunasekera, A., Zhang, X., Kunkel, T. A., Ebright, R. H. & Berman, H. M. (2001). Indirect readout of DNA sequence at the primary-kink site in the CAP–DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. *J. Mol. Biol.* **314**, 75–82.
37. Lopata, M., Schlieper, D., von Wilcken-Bergmann, B. & Muller-Hill, B. (1997). A lethal mutant of the catabolite gene activator protein CAP of *Escherichia coli*. *Biol. Chem.* **378**, 1153–1162.
38. Gunasekera, A., Ebright, Y. W. & Ebright, R. H. (1990). DNA-sequence recognition by CAP: role of the adenine N6 atom of base pair 6 of the DNA site. *Nucl. Acids Res.* **18**, 6853–6856.
39. Gunasekera, A., Ebright, Y. W. & Ebright, R. H. (1992). DNA sequence determinants for binding of the *Escherichia coli* catabolite gene activator protein. *J. Biol. Chem.* **267**, 14713–14720.
40. Zhang, X. P. & Ebright, R. H. (1990). Identification of a contact between arginine-180 of the catabolite gene activator protein (CAP) and base pair 5 of the DNA site in the CAP–DNA complex. *Proc. Natl Acad. Sci. USA*, **87**, 4717–4721.
41. Orengo, C. A. & Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617–635.
42. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
43. Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.
44. Orengo, C. A. (1999). CORA-topological fingerprints for protein structural families. *Protein Sci.* **8**, 699–715.
45. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282.

Edited by F. E. Cohen

(Received 31 October 2001; received in revised form 7 June 2002; accepted 7 June 2002)