

Supplementary material for

**Protein-DNA interactions: amino acid conservation
and the effects of mutations on binding specificity**

Nicholas M Luscombe and Janet M Thornton

Note to referees:

Data presented in the appendix will be presented on the web as supplementary material.

Appendix 1

Here we provide a detailed analysis of amino acid-base interactions in each of the individual 21 families, the importance of the interactions in binding and specificity, and the effect that amino acid mutations have on sequence recognition.

Mutation tables

Supplementary Tables A1.1-A1.22 summarise the variations of amino acid sequences at base-contacting positions in each protein family. Each table has two sections. The first lists the representative PDB structure, the number of aligned sequences, and number of organisms from which the sequences are found in. The biological or non-biological target DNA sequence is given and the bases in contact with the protein are underlined.

The second half of the table lists the aligned sequences, including its PDB or SWISS-PROT identity, full protein name and the organism from which it originates. We list all the base-contacting positions in the alignment and residues are numbered according to the PDB residue number in the representative structure. Positions that interact with bases in the target sequence are underlined, and those that interact according to the universal amino acid-base pairing (see below) are asterisked. At each position, we give the amino acid (and its universal pairing set) that is found in each protein sequence.

Families are divided into subfamilies according to the amino acid sequences at target-contacting positions. Proteins in the same subfamily have identical or similar (as defined by the rules of amino acid-base pairing) sets of amino acids and are expected to recognise the same DNA sequences. Proteins in different subfamilies have significant mutations at one or more positions and they are predicted to bind different DNA sequences. Mutations at positions that interact with bases outside the targets sequence are not considered for these purposes. The “individual proteins” category lists proteins with unique sets of target-contacting positions.

Finally, each table is accompanied by a brief description of the family and the alignment.

Amino acid sets

We previously classified amino acid-base interactions in 129 PDB structures according to their pairing preferences¹. We summarise our findings in Table A1.1, which highlights the interaction pairs that are universally favoured across all protein-DNA complexes. Interactions that follow the pairing described in the table are termed “universal” and those that do not, but nonetheless provide sequence specificity are

termed “context-dependent”. Sets (a)-(e) comprise amino acids that mainly bind bases using hydrogen bonds; the main interactions are bidentate (or bifurcated) bonds with single bases, or complex bonds that contact two or more adjacent base-steps. Arginine, lysine and histidine (sets a-b) favour guanine. Using their two side-chain donor atoms (double-donor) arginine and lysine can also interact with bases from adjacent base-steps if two acceptor atoms are presented (complex bonds). Histidine, however, are rarely involved in complex bonds. Likewise, serine of set (c) also binds guanine, but the hydroxyl acts as an acceptor+donor group for complex bonds. In set (d), asparagine and glutamine favour adenine and have acceptor+donor side chains. Although aspartate and glutamate of set (e) do not readily interact with DNA, the double-acceptor side chains have significantly different hydrogen-bonding properties to the other amino acids. All four sets are also expected to make water-mediated bonds.

Sets (f)-(h) combine amino acids that predominantly bind bases using van der Waals contacts. Phenylalanine and proline of set (f) favour adenine and thymine and produce ring-stacking interactions. Although threonine in set (g) has comparable hydrogen-bonding properties to serine, we observed that recognition of thymine through methyl-methyl contacts is more common than interactions using the hydroxyl group. Set (h) consists of small hydrophobic amino acids and tyrosine, which contacts bases indiscriminately. Finally, members of set (i) do not interact with bases in significant numbers.

Amino acid sets and the effect of mutations

The effects of amino acid mutations on DNA-binding specificity are best understood with these preferences in mind. Substitutions between amino acids in the same set are not expected to alter specificity (*eg* arginine to lysine). However, substitutions across different sets are likely to affect binding, although the size of the effect depends on the nature of the interaction. Amino acid-base preferences are much stronger and more specific for hydrogen bonds than van der Waals contacts. Thus, substitutions of hydrogen-bonding amino acids (sets a-e) will cause greater changes in specificity than substitutions between those that use van der Waals (sets f-i). Obviously, the anticipated effects are most likely to be correct if the interaction follows the universal pattern.

In total (excluding non-specific families), interactions from 43 target-contacting positions follow the universal amino acid-base pairings, while 25 positions produce context-dependent interactions. Typically, each family produces 2-3 universal interactions, and 1-2 non-universal interactions.

The most frequent hydrogen bonds are arginine, lysine or histidine (set a) with guanine. Complex bonds with multiple base-steps are also observed. For example, amide residues (set c) contact -TC- (leucine zipper family), -AC- (papillomavirus-1 E2 family) and -TT- (TBP family). Lysine (set a) and glutamate (set d) bind -GG- (papillomavirus-1 E2 family) and -CC- (RHR family) respectively. The most common van der Waals contacts are phenylalanine or proline with adenine or thymine (T-domain and TBP families).

The most common non-universal interactions use the peptide backbone (Prd paired domain, Trp repressor, and T-domain families). Also observed are van der Waals contacts from hydrophobic residues (set g and h) that do not usually provide specificity.

Table A1.1

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]	-	-

Table A1.2 Integration host factor family (Non-specific)

The integration host factor family contains architectural proteins that assist formation of high-order protein-DNA complexes such as those found in replication and long distance transcription regulation. Although sequence-dependent binding to a 35 base-pair site is recorded, very little consensus is observed between the sites, except for a poly-adenine tract².

Eight interacting positions are found, many of which are well conserved. Only the interaction from position 65 is of real importance in binding; the proline residue at this position intercalates between base-pairs to bend the DNA. The remaining positions effectively provide stabilising interactions for the complex.

47 and 66 are the most variable positions. 47 is outside the β -hairpin DNA-binding motif and consists of arginine that binds thymine in *lihfA*, serine, or hydrophobic amino acids. Lysine or glutamine at position 66 are involved in van der Waals contacts to multiple bases.

Table A1.2

Integration host factor family (Non-specific)										
Key base-contacting positions:		65 (base-pair intercalation)								
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1ihf chain A)							
			47	60	62	63*	64	65*	66	73
HLIK_ASFB7	Histone-like protein homolog	<i>African swine fever virus</i>	h (I)	a (R)	h (G)	b (H)	d (N)	f (P)	h (A)	h (I)
DBH_ANASP	DNA-binding protein Hu	<i>Anabaena sp</i>	h (V)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (I)
1hueA	Histone like protein	<i>Bacillus stearothermophilus</i>	h (I)	a (R)	h (G)	a (R)	d (N)	f (P)	d (Q)	h (I)
1wtuA	Transcription factor 1	<i>Bacillus subtilis</i>	g (T)	a (R)	h (G)	f (F)	d (N)	f (P)	d (Q)	h (I)
DBH_BACSU	DNA-binding protein II (Hb) (Hu)	<i>Bacillus subtilis</i>	h (I)	a (R)	h (G)	a (R)	d (N)	f (P)	d (Q)	h (I)
DBH_BORAD	DNA-binding protein Hbbu	<i>Borrelia andersonii</i>	a (R)	a (R)	h (A)	a (R)	d (N)	f (P)	d (Q)	h (V)
DBH_BORPR	DNA-binding protein Hbbu	<i>Borrelia parkeri</i>	a (R)	a (R)	h (A)	a (R)	d (N)	f (P)	d (Q)	h (V)
IHFB_BUCA	Integration host factor-β	<i>Buchnera aphidicola</i>	a (R)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (L)
DBH_CLOPA	DNA-binding protein Hu	<i>Clostridium pasteurianum</i>	h (V)	a (R)	h (G)	a (R)	d (N)	f (P)	a (R)	h (I)
1ihfA	Integration host factor	<i>Escherichia coli</i>	c (S)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (I)
1ihfB	Integration host factor	<i>Escherichia coli</i>	a (R)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (L)
DBHA_ECOL	DNA-binding protein Hu-a (Hu-2)	<i>Escherichia coli</i>	h (V)	a (R)	h (G)	a (R)	d (N)	f (P)	d (Q)	h (I)
DBHB_ECOL	DNA-binding protein Hu-b (Hu-1)	<i>Escherichia coli</i>	h (V)	a (R)	h (G)	a (R)	d (N)	f (P)	d (Q)	h (I)
DBH_GUITH	DNA-binding protein Hu homolog	<i>Guillardia theta</i>	h (V)	a (R)	h (G)	a (R)	d (N)	f (P)	a (R)	h (L)
DBHA_HAEI	Integration host factor-α	<i>Haemophilus influenzae</i>	h (I)	a (R)	h (G)	a (R)	d (N)	f (P)	d (Q)	h (I)
IHFA_HAEI	Integration host factor-α	<i>Haemophilus influenzae</i>	c (S)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (V)
IHFB_HAEI	Integration host factor-β	<i>Haemophilus influenzae</i>	a (R)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (L)
DBH_HELPY	DNA-binding protein Hu	<i>Helicobacter pylori</i>	h (I)	a (K)	h (G)	a (K)	h (V)	f (P)	h (G)	g (T)
IHFA_PASH	Integration host factor-α	<i>Pasteurella haemolytica</i>	c (S)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (V)
IHFB_PASH	Integration host factor-β	<i>Pasteurella haemolytica</i>	a (R)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (L)
DBH_PSEAE	DNA-binding protein Hu	<i>Pseudomonas aeruginosa</i>	h (V)	a (R)	h (G)	a (R)	d (N)	f (P)	d (Q)	h (I)
IHFA_PSEA	Integration host factor-α	<i>Pseudomonas aeruginosa</i>	c (S)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (I)
IHFB_PSEA	Integration host factor-β	<i>Pseudomonas aeruginosa</i>	a (R)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (L)
IHFB_PSEP	Integration host factor-β	<i>Pseudomonas putida</i>	a (R)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (L)
DBH1_RHIL	DNA-binding protein Hrl18	<i>Rhizobium leguminosarum</i>	h (A)	c (S)	h (G)	a (R)	d (N)	f (P)	c (S)	h (I)
DBH5_RHIL	DNA-binding protein Hrl53	<i>Rhizobium leguminosarum</i>	h (A)	g (T)	h (G)	a (R)	d (N)	f (P)	c (S)	h (I)
IHFA_RHOC	Integration host factor-α	<i>Rhodobacter capsulatus</i>	c (S)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (I)
IHFB_RHOC	Integration host factor-β	<i>Rhodobacter capsulatus</i>	a (R)	a (R)	h (G)	a (R)	d (N)	f (P)	a (R)	h (V)
DBH_STRTR	DNA-binding protein Hu	<i>Streptococcus thermophilus</i>	h (I)	a (R)	h (G)	a (R)	d (N)	f (P)	d (Q)	h (I)
DBH_STRLI	DNA-binding protein Hu (Hs1)	<i>Streptomyces lividans</i>	f (P)	a (R)	h (A)	a (R)	d (N)	f (P)	d (Q)	h (I)
DBH_SYNY3	DNA-binding protein Hu	<i>Synechocystis sp</i>	h (V)	a (R)	h (G)	a (R)	d (N)	f (P)	a (K)	h (I)
DBH_THEAC	DNA-binding protein Hta	<i>Thermoplasma acidophilum</i>	h (A)	a (R)	h (A)	a (R)	d (N)	f (P)	d (Q)	h (V)
DBH_THEMA	DNA-binding protein Hu	<i>Thermotoga maritima</i>	h (V)	a (R)	h (G)	h (V)	d (N)	f (P)	d (Q)	h (I)
DBH_THETH	DNA-binding protein II	<i>Thermus aquaticus</i>	g (T)	a (R)	h (G)	h (V)	a (K)	f (P)	h (G)	h (I)
Conservation score			57.4	82.8	76.7	72.6	79.6	100	56.2	70.2

Table A1.3. Dnase I family (Non-specific)

Dnase I proteins are non-specific endonucleases that degrade double-stranded DNA through hydrolysis of the phosphate backbone. Three positions interact with bases. Position 9 is either arginine or glutamine. Position 41 is a conserved arginine or lysine. Position 75 is mainly serine or threonine. None of the interactions follow the universal pairings. However, van der Waals contacts from position 9 and hydrogen bonds from positions 41 and 75 aid structural deformation of the DNA.

Table A1.3

Dnase I family (Non-specific)						
Key base-contacting positions:						
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 2dnj chain A)			
			9	41	75	
1dnkA	Deoxyribonuclease I	<i>Bos taurus</i>	a (R)	a (R)	c (S)	
2dnjA	Deoxyribonuclease I	<i>Bos taurus</i>	a (R)	a (R)	c (S)	
DRN1_BOVIN	Deoxyribonuclease I	<i>Bos taurus</i>	a (R)	a (R)	c (S)	
DRN1_MOUSE	Deoxyribonuclease I precursor	<i>Mus musculus</i>	a (R)	a (R)	c (S)	
DHP2_HUMAN	Dnase I homologue Dhp2 precursor	<i>Homo sapiens</i>	a (R)	a (K)	g (T)	
DRN1_HUMAN	Deoxyribonuclease I precursor	<i>Homo sapiens</i>	d (Q)	a (R)	c (S)	
DHP1_HUMAN	Dnase I homologue Dhp2 precursor	<i>Homo sapiens</i>	d (Q)	a (R)	d (Q)	
DRNL_HUMAN	Muscle-specific dnase I-like precursor	<i>Homo sapiens</i>	d (Q)	h (V)	g (T)	
DRN1_PIG	Deoxyribonuclease I	<i>Sus scrofa</i>	a (R)	a (R)	g (T)	
Conservation score			77.5	79.9	69	

Table A1.4. Taq polymerase family (Non-specific)

Taq polymerase catalyses the sequential addition of nucleotides during DNA replication. The protein is structurally almost identical to polymerase I.

Although the six base-contacting positions are not involved in recognition of particular DNA sequences, the interactions provided by them are important for accurate functioning of polymerase activity. The polar amino acids at positions 746 and 754 bind the O2 and N3 acceptor atoms on the minor groove edge of the base-pair in the active site. The two atoms occupy identical positions in all four Watson-Crick base-pairs, but not otherwise. Therefore, the residues provide a sequence independent method of checking the integrity of the nucleotide-pair being added to the DNA. Aromatic amino acids at positions 667 and 671 stabilise newly added bases through ring-stacking interactions and the polar amino acid at position 583 maintains the widened conformation of the minor groove.

The five important base-contacting positions are well conserved across all proteins from different species.

Table A1.4

Polymerase- β family (Non-specific)									
Key base-contacting positions:		234 (widing of DNA minor groove) 271, 280 (sequence-independent check of base-pair incorporation)							
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1bpy chain A)						
			34	35	38	234	271	279	280
TDT_BOVIN	DNA nucleotidylexotransferase- β	<i>Bos taurus</i>	f (F)	- (-)	g (T)	i (C)	h (G)	e (E)	a (R)
2bpfA	DNA polymerase- β	<i>Escherichia coli</i>	b (H)	a (K)	h (A)	a (K)	h (Y)	d (N)	a (K)
TDT_CHICK	DNA nucleotidylexotransferase- β	<i>Gallus gallus</i>	f (F)	- (-)	e (E)	c (S)	h (G)	h (G)	a (R)
1bpxA	DNA polymerase- β	<i>Homo sapiens</i>	b (H)	a (K)	h (A)	a (K)	h (Y)	d (N)	a (K)
1bpyA	DNA polymerase- β	<i>Homo sapiens</i>	b (H)	a (K)	h (A)	a (K)	h (Y)	d (N)	a (K)
1zqaA	DNA polymerase- β	<i>Homo sapiens</i>	b (H)	a (K)	h (A)	a (K)	h (Y)	d (N)	a (K)
DPOB_HUMAN	DNA polymerase- β	<i>Homo sapiens</i>	b (H)	a (K)	h (A)	a (K)	h (Y)	d (N)	a (K)
TDT_HUMAN	DNA nucleotidylexotransferase- β	<i>Homo sapiens</i>	f (F)	- (-)	g (T)	i (C)	h (G)	e (E)	a (R)
TDT_MONDO	DNA nucleotidylexotransferase- β	<i>Monodelphis domestica</i>	f (F)	- (-)	g (T)	a (K)	h (G)	e (E)	a (R)
TDT_MOUSE	DNA nucleotidylexotransferase- β	<i>Mus musculus</i>	f (F)	- (-)	h (A)	a (K)	h (G)	e (E)	a (R)
TDT_XENLA	DNA nucleotidylexotransferase- β	<i>Xenopus laevis</i>	f (F)	h (I)	- (-)	h (I)	h (G)	e (E)	a (R)
Conservation score			65.3	18.2	61.4	64.8	60.2	64.5	81.4

Table A1.5. Polymerase- β family (Non-specific)

Polymerase- β is a repair enzyme that restores a single nucleotide gap in double-stranded DNA, once the damaged or mismatched base is removed by excision enzymes.

Amino acids hold similar functions to those in Taq polymerase. Positions 271 and 280 are thought to check the integrity of the base-pair in the newly added nucleotide. Position 271 comprises tyrosine or glycine – although it is unclear how the latter achieves the equivalent interaction – and position 280 consists of a conserved arginine or lysine. A partly conserved lysine at position 234 helps widen the DNA minor groove.

Table A1.5

Polymerase Taq family (Non-specific)								
Key base-contacting positions:		583	(widening of minor groove)					
		667, 671	(stabilisation of newly added bases by ring-stacking)					
		746, 754	(sequence-independent check of base-pair incorporation)					
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1tau chain A)					
			583	667	671	677	746	754
DPO1_ANATH	DNA polymerase I	<i>Anaerocellum thermophilum</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPO1_AQUAE	DNA polymerase I	<i>Aquifex aeolicus</i>	d (N)	f (F)	h (Y)	h (G)	- (-)	d (Q)
DPO1_BACCA	DNA polymerase I	<i>Bacillus caldotenax</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPO1_BACST	DNA polymerase I	<i>Bacillus stearothermophilus</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPO1_BACSU	DNA polymerase I	<i>Bacillus subtilis</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPOL_BPSP1	DNA polymerase I	<i>Bacteriophage Sp01</i>	d (Q)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPOL_BPT3	DNA polymerase I	<i>Bacteriophage T3</i>	d (Q)	h (Y)	h (Y)	a (K)	a (R)	d (Q)
DPOL_BPT5	DNA polymerase I	<i>Bacteriophage T5</i>	d (Q)	f (F)	h (Y)	a (K)	a (R)	d (Q)
DPO1_CHLAU	DNA polymerase I	<i>Chloroflexus aurantiacus</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPO1_DEIRA	DNA polymerase I	<i>Deinococcus radiodurans</i>	d (N)	f (F)	h (Y)	a (R)	a (R)	d (Q)
DPO1_ECOLI	DNA polymerase I	<i>Escherichia coli</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPO1_HAEIN	DNA polymerase I	<i>Haemophilus influenzae</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPO1_HELPY	DNA polymerase I	<i>Helicobacter pylori</i>	d (N)	f (F)	h (Y)	a (K)	a (R)	d (Q)
DPO1_LACLC	DNA polymerase I	<i>Lactococcus lactis</i>	d (N)	f (F)	h (Y)	h (G)	a (R)	d (Q)
DPO1_MYCLE	DNA polymerase I	<i>Mycobacterium leprae</i>	d (N)	h (Y)	h (Y)	h (G)	a (R)	d (Q)
DPO1_MYCTU	DNA polymerase I	<i>Mycobacterium tuberculosis</i>	d (N)	h (Y)	h (Y)	h (G)	a (R)	d (Q)
DPOL_BPMD2	DNA polymerase I	<i>Mycobacteriophage D29</i>	g (T)	f (F)	h (Y)	h (G)	- (-)	d (Q)
DPOL_BPML5	DNA polymerase I	<i>Mycobacteriophage L5</i>	- (-)	f (F)	h (Y)	h (G)	- (-)	d (Q)
DPO1_RICPR	DNA polymerase I	<i>Rickettsia prowazekii</i>	d (N)	f (F)	h (Y)	h (A)	a (R)	d (Q)
DPO1_STRPN	DNA polymerase I	<i>Streptococcus pneumoniae</i>	d (N)	f (F)	h (Y)	h (G)	h (G)	d (Q)
1tau (chain A)	DNA polymerase Taq	<i>Thermus aquaticus</i>	d (N)	f (F)	h (Y)	a (R)	a (R)	d (Q)
DPO1_THECA	DNA polymerase I	<i>Thermus aquaticus</i>	d (N)	f (F)	h (Y)	a (R)	a (R)	d (Q)
DPO1_THEFL	DNA polymerase I	<i>Thermus aquaticus</i>	d (N)	f (F)	h (Y)	a (R)	a (R)	d (Q)
DPO1_THEFI	DNA polymerase I	<i>Thermus filiformis</i>	d (N)	f (F)	h (Y)	a (R)	a (R)	d (Q)
DPO1_TREPA	DNA polymerase I	<i>Treponema pallidum</i>	d (N)	f (F)	h (Y)	a (R)	a (R)	d (Q)
Conservation score			84.3	97.5	100	62.8	77.6	100

Table A1.6. Pu1 ETS domain family (Highly specific)

Members of the Pu1 ETS domain family share a conserved ETS domain that binds DNA mainly through a helix-turn-helix motif. Proteins are involved in regulation of expression of different genes during growth and development.

Two position from the probe α -helix contact the target sequence; positions 232 and 235 comprise arginines that interact with guanine in bidentate interactions (-..G...- and -.G....- respectively). NMR studies of have shown that equivalent amino acids in the human FLI-1 protein (FLI1_HUMAN) are also interacting ³.

The two amino acids are absolutely conserved for all family members and experimental evidence shows that point mutations at either alignment position abolishes DNA-binding activity ⁴. Although there are currently no diseases associated with point mutations in ETS domain proteins, disruption of the Pu1 genes in mice causes abnormalities in bone formation and the immune system in mice ⁵.

Table A1.6

Pu1 ETS domain family (Highly specific)						
Representative structure:		1pue chain E				
Helix-turn-helix motif:		Residues 208-240				
Target-contacting positions:		232 (-.G..-) 235 (-.G...-)				
Target sequence:		-(C/A)GGA(AT)-				
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1pue chain E)			
			228	232*	235*	
E74A_DROME	E74A	<i>Drosophila melanogaster</i>	e (E)	a (R)	a (R)	
ETS3_DROME	D-ETS-3	<i>Drosophila melanogaster</i>	e (D)	a (R)	a (R)	
ETS4_DROME	D-ETS-4	<i>Drosophila melanogaster</i>	e (D)	a (R)	a (R)	
ETS6_DROME	D-ETS-6	<i>Drosophila melanogaster</i>	e (D)	a (R)	a (R)	
PNT1_DROME	D-ETS-2	<i>Drosophila melanogaster</i>	e (E)	a (R)	a (R)	
POK_DROME	Pokurri	<i>Drosophila melanogaster</i>	e (D)	a (R)	a (R)	
1bc7C	SAP-1	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
1sttA	ETS-1	<i>Homo sapiens</i>	e (E)	a (R)	a (R)	
ELF1_HUMAN	ELF-1	<i>Homo sapiens</i>	e (E)	a (R)	a (R)	
ELK1_HUMAN	ELK-1	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
ELK3_HUMAN	ELK-3	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
ERF_HUMAN	ERF	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
FLI1_HUMAN	FLI-1	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
ETV1_HUMAN	ETV-1	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
ETV2_HUMAN	ETV-2	<i>Homo sapiens</i>	e (E)	a (R)	a (R)	
ETV3_HUMAN	ETV-3	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
ETV6_HUMAN	ETV-6	<i>Homo sapiens</i>	e (E)	a (R)	a (R)	
PU1_HUMAN	Pu1	<i>Homo sapiens</i>	d (Q)	a (R)	a (R)	
SAPA_HUMAN	SAP-1A	<i>Homo sapiens</i>	e (D)	a (R)	a (R)	
ERG_LYTVA	ERG	<i>Lytechinus variegatus</i>	e (D)	a (R)	a (R)	
ETS2_LYTVA	ETS-2	<i>Lytechinus variegatus</i>	e (E)	a (R)	a (R)	
1pue	Pu1-ETS	<i>Mus musculus</i>	e (E)	a (R)	a (R)	
1etc	ETS-1	<i>Mus musculus</i>	e (E)	a (R)	a (R)	
MYBE_AVILE	P135-Gag-Myb-ETS transforming protein	<i>Avian leukemia virus E26</i>	e (E)	a (R)	a (R)	
ERG_CHICK	ERG	<i>Gallus gallus</i>	e (D)	a (R)	a (R)	
ET1A_XENLA	C-ETS-1A	<i>Xenopus laevis</i>	e (E)	a (R)	a (R)	
Conservation score			80.6	100	100	

Table A1.7. Prd paired domain family (Highly specific)

The paired domain family comprises a set of helix-turn-helix transcription regulators that are important for developmental processes ⁶.

There are six target-contacting positions, the first five being part of the winged helix-turn-helix motif. Positions 14 and 15 are part of the β -hairpin in the motif and contact the DNA minor groove: both act as acceptors from guanine N2 atoms (-.....G.....- and -.....G.....- respectively). Positions 47-49 are part of the recognition helix. At position 47, histidine in IpdnC recognises guanine (-...G.....-). Positions 48 and 49 make van der Waals contacts with thymines (-...T.T'.....- and -.....T'.....- respectively). Finally, the peptide backbone of position 70, part an interdomain loop, acts as donor to the guanine N3 atom in the minor groove.

Amino acid substitutions are only found at position 47, where asparagine replaces histidine in the Pax-6 (PAX6_BRARE) and Pax-4 proteins (PAX4_MOUSE and PAX4_HUMAN). The mutation is accompanied by a change in the corresponding target sequence from guanine to adenine. However, as the targets are non-biological, the *in vivo* effect of the mutation is unclear. Biochemical studies also suggest that the less well-conserved C-terminal helix-turn-helix domain may play a role in DNA-binding in proteins such as Pax-5 and Pax-6 ⁷.

Point mutations in the N-terminal helix-turn-helix motif disrupt DNA-binding and are associated with developmental abnormalities in mice and human. In particular, mutations at position 15 (G \rightarrow S) of Pax-1 results in malformation of the mouse vertebral column ⁸, and mutations at positions 14 (N \rightarrow H) and 48 (G \rightarrow A) of Pax-3 causes Waardenburg's syndrome, which manifests itself in deafness and pigmentary disorders ⁹⁻¹².

Point mutations at backbone-contacting positions in Pax-6 (positions 17, 23 and 63) and frameshift or deletion mutations also result in Waardenburg's syndrome and related disorders ¹³. All point mutations are found in the N-terminal domain, highlighting its importance in binding and more disorders are expected from mutations in other family members.

Table A1.7

Prd paired domain family (Highly specific)										
Representative structure:	1pdn chain C									
Helix-turn-helix motif:	Residues 36-60									
Target-contacting positions:	14 (-.....G.....-) 15 (-.....G.....-) 47* (-...G.....-) 48 (-....T.T.....-) 49 (-.....T.....-) 70 (-.....G..-)									
Target sequence:	-AACGTCACGGTTGAC- (1pdn chainC) -nCGTCACG(G/C)TT(G/C)Pu- (Prd) -nTnGTCAPyGCPuTGA- (Pax-2) -Py(G/C)GTPy(A/C)CGCnnCANtGnnPy- (Pax-5) -AnnTTCACGC(A/T)T(G/C)AnT(G/T)(A/C)nPy- (Pax-6)									
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1pdn chain C)							
			14	15	47*	48	49	68	69	70
Subfamily 1										
1pdnC	Paired box protein Prd	<i>Drosophila melanogaster</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
GSBD_DROME	Gooseberry distal protein (Bsh9)	<i>Drosophila melanogaster</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
GSBP_DROME	Gooseberry proximal protein (Bsh4)	<i>Drosophila melanogaster</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
POXM_DROME	Paired box pox-meso protein	<i>Drosophila melanogaster</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
POXN_DROME	Paired box pox-neuro protein	<i>Drosophila melanogaster</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX1_HUMAN	Paired box protein Pax-1	<i>Homo sapiens</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX2_HUMAN	Paired box protein Pax-2	<i>Homo sapiens</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX3_HUMAN	Paired box protein Pax-3 (Hup2)	<i>Homo sapiens</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX7_HUMAN	Paired box protein Pax-7 (Hup1)	<i>Homo sapiens</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX9_HUMAN	Paired box protein Pax-9	<i>Homo sapiens</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX7_MOUSE	Paired box protein Pax-7	<i>Mus musculus</i>	d (N)	g (G)	b (H)	g (G)	i (C)	g (I)	g (G)	g (G)
Subfamily 2										
PAX6_HUMAN	Paired box protein Pax-6	<i>Homo sapiens</i>	d (N)	g (G)	d (N)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX4_HUMAN	Paired box protein Pax-4	<i>Homo sapiens</i>	d (N)	g (G)	d (N)	g (G)	i (C)	g (I)	g (G)	g (G)
PAX4_MOUSE	Paired box protein Pax-4	<i>Mus musculus</i>	d (N)	g (G)	d (N)	g (G)	i (C)	h (I)	h (G)	h (G)
Conservation score			100	100	85.5	100	100	100	100	100

Table A1.8. Trp repressor family (Highly specific)

The members of the Trp repressor family are involved in regulation of tryptophan synthesis and bind DNA using a helix-turn-helix motif.

Three positions in the motif contact the target sequence. Arginine at position 69 hydrogen bonds with the N7 atom of guanine or adenine, therefore differentiating between purine and pyrimidine bases. Alanine at position 80 recognises thymine (-T'...-) via methyl-methyl contacts. Finally, positions 79 and 80 recognise the first two bases of the target sequence (-AG..-) with a network of water-mediated bonds from the peptide backbone ¹⁴. These water-mediated interactions are thought to be especially important for target site recognition ¹⁵.

All target-contacting positions are well conserved across species and the observed mutation at position 79 is not expected to affect binding. Mutation analysis demonstrates that all three positions are required for specificity and selective amino acid substitutions alter recognition of the DNA sequence ^{16,17}. No diseases are yet associated with point mutations in the protein.

Table A1.8

Trp repressor family (Highly specific)						
Representative structure:		1trr chain A				
Helix-turn-helix motif:		Residues 68-91				
Target-contacting positions:		69* (-...Pu-)				
		79 (AG...)				
		80 (-T'...)				
Target sequence:		-AGPuPu-				
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1trr chain A)			
			69*	79	80	
1trrA	Trp operon repressor	<i>Escherichia coli</i>	a (R)	h (I)	h (A)	
2wrpR	Trp operon repressor	<i>Escherichia coli</i>	a (R)	h (I)	h (A)	
3wrp	Trp operon repressor	<i>Escherichia coli</i>	a (R)	h (I)	h (A)	
1troC	Trp operon repressor	<i>Escherichia coli</i>	a (R)	h (I)	h (A)	
1troE	Trp operon repressor	<i>Escherichia coli</i>	a (R)	h (I)	h (A)	
TRPR_ENTAE	Trp operon repressor	<i>Enterobacter aerogenes</i>	a (R)	h (I)	h (A)	
TRPR_ENTCL	Trp operon repressor	<i>Enterobacter cloacae</i>	a (R)	h (I)	h (A)	
TRPR_HAEIN	Trp operon repressor	<i>Haemophilus influenzae</i>	a (R)	h (A)	h (A)	
Conservation score			100	86.3	100	

Table A1.9. Loop-sheet-helix family (p53) (Highly specific)

The loop-sheet-helix family comprises p53 proteins with cancer-inhibiting properties. It functions as a transcription factor with roles in controls of cell cycle progression and apoptosis and binds DNA through a loop-sheet-helix zinc-coordinating motif.

Three positions interact with the target site. Positions 120 and 280 comprise basic residues that bind guanine (-.G...- and -...G'.- respectively). Position 277 has a cysteine which contacts cytosine (-..C'..-). Variations in the target sequence suggest that lysine at position 120 may also recognise adenine, but guanine is predicted to be preferred because of the possibility of a bidentate interaction. Cysteine at position 277 could also recognise thymine by using the side chain as a hydrogen bond donor rather than acceptor.

Base-contacting positions are very highly conserved across proteins from the represented species. The only amino acid mutation is at position 120 where a gap is present in the N-terminal end of a sequence fragment (P53_EQUAS). Numerous studies have linked over 250 disease-related point mutations in the p53 sequence to many forms of cancer. Among them are mutations at the target-contacting positions which are expected to disrupt recognition of the target sequence severely^{18,19}.

Table A1.9

Loop-sheet-helix (p53) family (Highly specific)						
Representative structure:		1tsr chain A				
Probe α -helix motif:		Residues 278-287				
Target-contacting positions:		120* (-.Pu...-)				
		277 (-..C'..-)				
		280* (-...G'..-)				
Target sequence:		-PuPuPuC(A/T)-				
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1tsr chain A)			
			120*	277	280*	
P53_BOVIN	P53 tumor suppressor	<i>Bos taurus and Bos indicus</i>	a (K)	i (C)	a (R)	
P53_BRARE	P53 tumor suppressor	<i>Brachydanio rerio</i>	a (K)	i (C)	a (R)	
P53_CANFA	P53 tumor suppressor (fragment)	<i>Canis familiaris</i>	a (K)	i (C)	a (R)	
P53_CRIGR	P53 tumor suppressor	<i>Cricetulus griseus</i>	a (K)	i (C)	a (R)	
P53_EQUAS	P53 tumor suppressor (fragment)	<i>Equus asinus</i>	- (-)	i (C)	a (R)	
P53_CHICK	P53 tumor suppressor	<i>Gallus gallus</i>	a (K)	i (C)	a (R)	
1tsrA	P53 tumor suppressor	<i>Homo sapiens</i>	a (K)	i (C)	a (R)	
1tsrB	P53 tumor suppressor	<i>Homo sapiens</i>	a (K)	i (C)	a (R)	
P53_MOUSE	P53 tumor suppressor	<i>Mus musculus</i>	a (K)	i (C)	a (R)	
P53_ORYLA	P53 tumor suppressor	<i>Oryzias latipes</i>	a (K)	i (C)	a (R)	
P53_PLAFE	P53 tumor suppressor	<i>Platichthys flesus</i>	a (K)	i (C)	a (R)	
P53_SALIR	P53 tumor suppressor	<i>Salmo irideus</i>	a (K)	i (C)	a (R)	
P53_XENLA	P53 tumor suppressor	<i>Xenopus laevis</i>	a (K)	i (C)	a (R)	
Conservation score			84.4	100	100	

Table A1.10. Leucine zipper family (Highly specific)

The leucine zipper family contains proteins that have the leucine zipper motif and binds as a dimer by symmetrically inserting long α -helix in the DNA major groove.

Two positions bind the target sequence. Asparagine at position 235 hydrogen bonds with thymine and cytosine that are diagonally positioned in a complex interaction (-CT'..- or -TC'..-). Arginine at position 243 preferentially binds guanine (-...G-) where available.

The two target-contacting positions are absolutely conserved for all members and therefore recognise both half-site sequences. Experimental evidence shows that the dimeric proteins differentiate between the full palindromic target sequences by recognising the spacing between half-sites, which is typically 0-2 base-pairs ²⁰. Amino acid mutations are found in the region that determines the positioning of the probe helices (residues 248--252) and the various combinations that can be achieved through heterodimer formation increase the number of possible target sites. However, in addition to their own target sites, many dimers bind others with high affinity: GCN4 and Fos/Jun proteins also bind the ATF/CREB sites and Zta to the Ap-1 site ²¹⁻²⁴.

Mutagenesis studies demonstrate that point mutations at the target-contacting positions alter recognition while those at non-target-contacting positions have little effect ²³. No diseases are yet associated with point mutations in the DNA-binding regions of the protein.

Table A1.10

Leucine zipper family (Highly specific)						
Representative structure:	2dgc chain A					
Probe α -helix motif:	Residues 230-250					
Target-contacting positions:	235* (-CT'..- or -TC'..-) 243* (-...n-)					
Target sequence:	-TGnn- -CAnn-					
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 2dgc chain A)			
			235*	238	239	243*
CREB_BOVIN	CREB2	<i>Bos taurus</i>	d (N)	h (A)	h (A)	a (R)
JUNB_CYPCA	Transcription factor Jun-B	<i>Cyprinus carpio</i>	d (N)	h (A)	h (A)	a (R)
AP1_DROME	Transcription factor AP-1	<i>Drosophila melanogaster</i>	d (N)	h (A)	h (A)	a (R)
AP1_CHICK	Transcription factor AP-1	<i>Gallus gallus</i>	d (N)	h (A)	h (A)	a (R)
JUND_CHICK	Transcription factor Jun-D	<i>Gallus gallus</i>	d (N)	h (A)	h (A)	a (R)
ATFA_HUMAN	Transcription factor Atf-A/Atf-A-de	<i>Homo sapiens</i>	d (N)	h (A)	h (A)	a (R)
ATF1_HUMAN	cAMP-dependent factor Atf-A	<i>Homo sapiens</i>	d (N)	h (A)	h (A)	a (R)
CREA_HUMAN	cAMP-responsive element modulator	<i>Homo sapiens</i>	d (N)	h (A)	h (A)	a (R)
JUNB_HUMAN	Transcription factor Jun-B (G0S3)	<i>Homo sapiens</i>	d (N)	h (A)	h (A)	a (R)
AP1_KLULA	AP-1-like transcription factor	<i>Kluyveromyces lactis</i>	d (N)	h (A)	d (Q)	a (R)
ATF2_MOUSE	Transcription factor Atf2 (Mxpb)	<i>Mus musculus</i>	d (N)	h (A)	h (A)	a (R)
CPC1_NEUCR	Cross-pathway control protein 1	<i>Neurospora crassa</i>	d (N)	h (A)	h (A)	a (R)
TAF1_TOBAC	Taf-1 (fragment)	<i>Nicotiana tabacum</i>	d (N)	c (S)	h (A)	a (R)
1dgcA	GCN4 protein	<i>Saccharomyces cerevisiae</i>	d (N)	h (A)	h (A)	a (R)
2dgcA	GCN4 protein	<i>Saccharomyces cerevisiae</i>	d (N)	h (A)	h (A)	a (R)
AP1_SCHPO	AP-1-like transcription factor	<i>S. pombe</i>	d (N)	h (A)	d (Q)	a (R)
HBPA_WHEAT	Transcription factor HBP-1a	<i>Triticum aestivum</i>	d (N)	c (S)	h (A)	a (R)
EMF1_WHEAT	DNA-binding EmBP-1 protein	<i>Triticum aestivum</i>	d (N)	c (S)	h (A)	a (R)
Conservation score			100	86.7	89.5	100

Table A1.11. Papillomavirus-1 E2 family (Highly specific)

The papillomavirus-E2 protein regulates transcription at all viral promoters and, in conjunction with the E1 protein, initiates DNA replication. DNA binding is mainly through a probe α -helix.

Two positions bind the target sequence: asparagine or glutamine at position 336 binds adenine and cytosine in a complex interaction (-AC....-) and arginine or lysine at position 339 binds two guanines (-.G'G'..-). The DNA is bent by 45° to allow the contacts to occur and suggests that there is indirect sequence recognition.

The amino acids at both target-contacting positions are highly conserved across the different viruses. Mutation analysis shows that the individual contributions from the amino acids are essential for specific binding²⁵.

Table A1.11

Papillomavirus-1 E2 family (Highly specific)						
Representative structure:		2bop chain A				
Probe α -helix motif:		Residues 335-348				
Target-contacting positions:		336* (-AC....-) 339* (-.G'G'..-)				
Target sequence:		-ACCnnn- (E2-BS promoter)				
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 2dgc chain A)			
			336*	339*	340	343
2bopA	Regulatory protein E2	<i>Bovine papillomavirus 1</i>	d (N)	a (K)	i (C)	f (F)
VE2_BPV1	Regulatory protein E2	<i>Bovine papillomavirus 1</i>	d (N)	a (K)	i (C)	f (F)
VE2_BPV4	Regulatory protein E2	<i>Bovine papillomavirus 4</i>	d (N)	a (K)	i (C)	a (R)
VE2_COPV	Regulatory protein E2	<i>Canine papillomavirus</i>	d (N)	a (K)	i (C)	h (Y)
VE2_PCPV1	Regulatory protein E2	<i>Chimpanzee papillomavirus 1</i>	d (N)	a (K)	i (C)	h (Y)
VE2_PAPVD	Regulatory protein E2	<i>Deer papillomavirus</i>	d (N)	a (K)	i (C)	f (F)
VE2_PAPVE	Regulatory protein E2	<i>Elk papillomavirus</i>	d (N)	a (K)	i (C)	f (F)
VE2_HPVO3	Regulatory protein E2	<i>Human papillomavirus 3</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO4	Regulatory protein E2	<i>Human papillomavirus 4</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO5	Regulatory protein E2	<i>Human papillomavirus 5</i>	d (N)	a (K)	d (N)	d (N)
VE2_HPVO6A	Regulatory protein E2	<i>Human papillomavirus 6a</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO7	Regulatory protein E2	<i>Human papillomavirus 7</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO8	Regulatory protein E2	<i>Human papillomavirus 8</i>	d (N)	a (K)	i (C)	d (N)
VE2_HPVO9	Regulatory protein E2	<i>Human papillomavirus 9</i>	d (N)	a (K)	i (C)	f (F)
VE2_HPVO10	Regulatory protein E2	<i>Human papillomavirus 10</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO11	Regulatory protein E2	<i>Human papillomavirus 11</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO12	Regulatory protein E2	<i>Human papillomavirus 12</i>	d (N)	a (K)	i (C)	d (N)
VE2_HPVO13	Regulatory protein E2	<i>Human papillomavirus 13</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO14	Regulatory protein E2	<i>Human papillomavirus 14</i>	d (N)	a (R)	i (C)	d (N)
VE2_HPVO15	Regulatory protein E2	<i>Human papillomavirus 15</i>	d (N)	a (K)	i (C)	f (F)
VE2_HPVO16	Regulatory protein E2	<i>Human papillomavirus 16</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO17	Regulatory protein E2	<i>Human papillomavirus 17</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO18	Regulatory protein E2	<i>Human papillomavirus 18</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO19	Regulatory protein E2	<i>Human papillomavirus 19</i>	d (N)	a (R)	c (S)	d (N)
VE2_HPVO1A	Regulatory protein E2	<i>Human papillomavirus 1a</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO20	Regulatory protein E2	<i>Human papillomavirus 20</i>	d (N)	a (K)	i (C)	d (N)
VE2_HPVO22	Regulatory protein E2	<i>Human papillomavirus 22</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO23	Regulatory protein E2	<i>Human papillomavirus 23</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO24	Regulatory protein E2	<i>Human papillomavirus 24</i>	d (N)	a (K)	i (C)	d (N)
VE2_HPVO26	Regulatory protein E2	<i>Human papillomavirus 26</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO27	Regulatory protein E2	<i>Human papillomavirus 27</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO28	Regulatory protein E2	<i>Human papillomavirus 28</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO29	Regulatory protein E2	<i>Human papillomavirus 29</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO30	Regulatory protein E2	<i>Human papillomavirus 30</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO31	Regulatory protein E2	<i>Human papillomavirus 31</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO32	Regulatory protein E2	<i>Human papillomavirus 32</i>	d (N)	a (K)	i (C)	i (W)
1dhmA	Regulatory protein E2	<i>Human papillomavirus 33</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO33	Regulatory protein E2	<i>Human papillomavirus 33</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO34	Regulatory protein E2	<i>Human papillomavirus 34</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO35	Regulatory protein E2	<i>Human papillomavirus 35</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO36	Regulatory protein E2	<i>Human papillomavirus 36</i>	d (N)	a (K)	i (C)	d (N)
VE2_HPVO37	Regulatory protein E2	<i>Human papillomavirus 37</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO38	Regulatory protein E2	<i>Human papillomavirus 38</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO39	Regulatory protein E2	<i>Human papillomavirus 39</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO40	Regulatory protein E2	<i>Human papillomavirus 40</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO41	Regulatory protein E2	<i>Human papillomavirus 41</i>	d (N)	a (R)	i (C)	h (Y)
VE2_HPVO42	Regulatory protein E2	<i>Human papillomavirus 42</i>	d (N)	a (K)	i (C)	f (F)
VE2_HPVO44	Regulatory protein E2	<i>Human papillomavirus 44</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO45	Regulatory protein E2	<i>Human papillomavirus 45</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO47	Regulatory protein E2	<i>Human papillomavirus 47</i>	d (N)	a (K)	i (C)	d (N)
VE2_HPVO48	Regulatory protein E2	<i>Human papillomavirus 48</i>	d (N)	a (K)	i (C)	d (N)
VE2_HPVO49	Regulatory protein E2	<i>Human papillomavirus 49</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO50	Regulatory protein E2	<i>Human papillomavirus 50</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO51	Regulatory protein E2	<i>Human papillomavirus 51</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO52	Regulatory protein E2	<i>Human papillomavirus 52</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO53	Regulatory protein E2	<i>Human papillomavirus 53</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO54	Regulatory protein E2	<i>Human papillomavirus 54</i>	d (N)	a (K)	i (C)	d (Q)
VE2_HPVO55	Regulatory protein E2	<i>Human papillomavirus 55</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO56	Regulatory protein E2	<i>Human papillomavirus 56</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO57	Regulatory protein E2	<i>Human papillomavirus 57</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO58	Regulatory protein E2	<i>Human papillomavirus 58</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO60	Regulatory protein E2	<i>Human papillomavirus 60</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO61	Regulatory protein E2	<i>Human papillomavirus 61</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO63	Regulatory protein E2	<i>Human papillomavirus 63</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO65	Regulatory protein E2	<i>Human papillomavirus 65</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO66	Regulatory protein E2	<i>Human papillomavirus 66</i>	d (N)	a (K)	i (C)	h (Y)
VE2_HPVO70	Regulatory protein E2	<i>Human papillomavirus 70</i>	d (N)	a (K)	i (C)	h (Y)
VE2_CRPVK	Regulatory protein E2	<i>Rabbit papillomavirus</i>	d (N)	a (K)	i (C)	h (Y)
VE2_RHPV1	Regulatory protein E2	<i>Rhesus papillomavirus 1</i>	d (N)	a (K)	i (C)	f (F)
Conservation score			100	97	96.5	77.6

Table A1.12. TATA box-binding family (Highly specific)

The TATA box-binding proteins form an essential component of the multiprotein transcription initiator complex that assembles on RNA polymerase promoters containing the sequence –TATA(T/A)A(T/A)n- (eukaryotes) or TTTA(T/A)Ann- (archaea). Binding is conducted through a pseudo-symmetric β -sheet structure in a widened DNA minor groove.

Twelve positions contact the target sequence. Two pairs of phenylalanine side chains (positions 99 and 116, 190 and 207) are intercalated between the first and last base-steps of the target, causing substantial kinking. Proline at position 191 interacts with adenine (-A'.....-), stabilising the kink formation and ensuring binding of the target sequence in the correct orientation by specifically adenine at the first base-step in the target²⁶. Asparagines at positions 69 and 159 are in complex interactions with base-pairs at the centre of the target site (-...T'(T'/A')...- and -...A(A/T)...- respectively). Threonine at position 215 acts as donor to adenine (-...A....-). Finally, hydrophobic amino acids at positions 71, 114, 161 and 205 interact with target bases via van der Waals contacts. Substitutions at positions 159 and 161 involve amino acids that are able to produce equivalent interactions. At position 215, the substituting serine does not interact in archaeal proteins (IaisA and TF2D_PYRKO); however, there is no decrease in specificity²⁷. Mutagenesis studies show that deletion of the intercalating phenylalanines abolishes binding. Replacement of the amide residues at positions 69 and 159 results in decreased specificity. Finally substitutions at positions 71, 114, 161 and 205 for larger or polar amino acids leads to loss of binding²⁸. No point mutations are associated with diseases, but as complex formation by TATA box-binding proteins is considered the nucleating event for transcription initiation, the organism is unlikely to survive with a dysfunctional protein.

In addition to the amino acid-base interactions described above, indirect read-out through DNA structure deformation and surface complementarity is believed to play a major role in binding. The proteins from thermophilic genomes are also thought to use bridging cations; the amino acids used for these interactions do not make direct contacts and therefore they are not seen in the table²⁹.

Table A1.12

TATA box-binding protein family (Highly specific)

Representative structure:	1ytb chain A														
Target-contacting positions:	69* (-...T(T/A)...) 71 (-...(T/A)...) 99* (-.....(T/A)n-) 114 (-.....A(T/A)-) 116* (-.....(T/A)n-) 159* (-...A(A/T)...) 161 (-...A(T)...) 190* (-T(A/T).....) 191* (A.....) 205 (-...T.....) 207* (-A(T/A).....) 215 (-...A.....)														
Target sequence:	-TAT(T/A)A(T/A)n- -TTTA(T/A)Ann-														
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1pue chain E)												
			69*	71	99*	114	116*	159*	161	162	190*	191*	205	207*	215
TF2D_ACACA	TFIID	<i>Acanthamoeba castellanii</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_ACECL	TFIID	<i>Acetabularia cliftonii</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF21_ARATH	TFIID-1	<i>Arabidopsis thaliana</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_CAEEL	TFIID	<i>Caenorhabditis elegans</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_CANAL	TFIID	<i>Candida albicans</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_DICDI	TFIID	<i>Dictyostelium discoideum</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_ENTHI	TFIID	<i>Entamoeba histolytica</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
1cdwA	TFIID	<i>Homo sapiens</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_METJA	TFIID	<i>Methanococcus jannaschii</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (Q)	i (M)	h (V)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_PYRKO	TFIID	<i>Pyrococcus kodakaraensis</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (Q)	i (M)	h (V)	f (F)	f (P)	h (L)	f (F)	c (S)
1aisA	TFIID	<i>Pyrococcus woesei</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (Q)	i (M)	h (V)	f (F)	f (P)	h (L)	f (F)	c (S)
1ytbA	TFIID	<i>Saccharomyces cerevisiae</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_SULSH	TFIID	<i>Sulfolobus shibatae</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (Q)	h (I)	h (V)	f (F)	f (P)	h (L)	f (F)	g (T)
TF2D_TETTH	TFIID	<i>Tetrahymena thermophila</i>	d (N)	h (V)	f (F)	h (L)	f (F)	d (N)	h (V)	h (G)	f (F)	f (P)	h (L)	f (F)	g (T)
Conservation score			100	100	100	100	100	80.2	82.2	75.8	100	100	100	100	88.1

Table A1.13. T-domain family (Highly specific)

The proteins in this family use a conserved T-domain to bind operons for genes that are essential in tissue specification, morphogenesis and organogenesis. Expression of the different proteins is tissue specific^{30,31}. DNA binding is through a helical bundle and β -hairpin.

Four positions interact with the target sequence. Position 67 has a conserved arginine that binds guanine in a bidentate interaction (-...G'....-). Position 211 contains a phenylalanine which uses the peptide backbone to specifically bind the guanine N2 atom (-.....G'..-) on the minor groove edge. Alanine at position 195 is in a methyl-methyl contact with thymine (-T.....-). Finally position 215 contains a phenylalanine that contacts adenine on the minor groove edge (-....A...-); steric effects with the side chain prevents binding to other bases because of atoms that extend out of the ring structure (N2 of guanine, O2 of pyrimidine).

There are single examples of mutations at positions 67, 195 and 215. In the first arginine is substituted by glycine in the mouse T-box 5 protein (TBX5_MOUSE), in the second alanine is substituted by serine in mouse T-box 4 and 5 and human T-box 5 proteins (TBX4_MOUSE, TBX5_MOUSE, TBX5_HUMAN), and in the third phenylalanine is substituted for leucine in the optomer-blind protein (OMB_MOUSE). The amino acid substitutions, especially those at positions 67 and 195, may alter recognition of the target sequence slightly; however, as the target sequence provided above is non-biological, the *in vivo* effect is unknown.

So far, no point mutations at the base-contacting positions have been linked to diseases. However, numerous stop-codon and missense mutations within the DNA-binding domain of the Tbx-5 protein are shown to cause Holt-Oram syndrome, in which there are developmental abnormalities in the heart and limbs. Other developmental disorders are expected to be linked with point mutations in members of the family^{32,33}.

Table A1.13

T-domain family (Highly specific)								
Representative structure:	1xbr chain A							
Target-contacting positions:	67* (-...G'....-) 195 (T.....-) 211 (-.....G'...-) 215* (-.....A...-)							
Target sequence:	-TnnCACCT-							
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1pue chain E)					
			67*	195	211	215*		
Subfamily 1								
BRA1_BRAFL	Ambra-1	<i>Branchiostoma floridae</i>	a (R)	h (A)	f (F)	f (F)		
BRA2_BRAFL	Ambra-2	<i>Branchiostoma floridae</i>	a (R)	h (A)	f (F)	f (F)		
BRAC_BRARE	Brachyury protein homolog	<i>Brachydanio rerio</i>	a (R)	h (A)	f (F)	f (F)		
TBX6_BRARE	TBX-6 (brain)	<i>Brachydanio rerio</i>	a (R)	h (A)	f (F)	f (F)		
TBX2_CAEEL	TBX-2	<i>Caenorhabditis elegans</i>	a (R)	h (A)	f (F)	f (F)		
TBX9_CAEEL	TBX-9	<i>Caenorhabditis elegans</i>	a (R)	h (A)	f (F)	f (F)		
TBX8_CAEEL	TBX-8	<i>Caenorhabditis elegans</i>	a (R)	h (A)	f (F)	f (F)		
TX12_CAEEL	TBX-12	<i>Caenorhabditis elegans</i>	a (R)	h (A)	f (F)	f (F)		
BYN_DROME	T-related protein	<i>Drosophila melanogaster</i>	a (R)	h (A)	f (F)	f (F)		
H15_DROME	TBX-H15	<i>Drosophila melanogaster</i>	a (R)	h (A)	f (F)	f (F)		
TBXL_CHICK	TBX-6L	<i>Gallus gallus</i>	a (R)	h (A)	f (F)	f (F)		
TBXT_CHICK	TBXT	<i>Gallus gallus</i>	a (R)	h (A)	f (F)	f (F)		
BRAC_HALRO	Brachyury protein homolog (As-t)	<i>Halocynthia roretzi</i>	a (R)	h (A)	f (F)	f (F)		
BRC2_HALRO	TBX-2 containing (As-T2)	<i>Halocynthia roretzi</i>	a (R)	h (A)	f (F)	f (F)		
BRAC_HEMPU	Brachyury protein homolog (hpta)	<i>Hemicentrotus pulcherrimus</i>	a (R)	h (A)	f (F)	f (F)		
TBR1_HUMAN	TBX-1 brain protein	<i>Homo sapiens</i>	a (R)	h (A)	f (F)	f (F)		
TBX2_HUMAN	TBX-2 (kidney, lung, placenta)	<i>Homo sapiens</i>	a (R)	h (A)	f (F)	f (F)		
TBX1_MOUSE	TBX-1 (testis)	<i>Mus musculus</i>	a (R)	h (A)	f (F)	f (F)		
TBX3_MOUSE	TBX-3 (lung, brain, heart)	<i>Mus musculus</i>	a (R)	h (A)	f (F)	f (F)		
TBX6_MOUSE	TBX-6 (neural)	<i>Mus musculus</i>	a (R)	h (A)	f (F)	f (F)		
1xbrA	T-box protein	<i>Xenopus laevis</i>	a (R)	h (A)	f (F)	f (F)		
EOMD_XENLA	Eomesodermin	<i>Xenopus laevis</i>	a (R)	h (A)	f (F)	f (F)		
VEGT_XENLA	T-box protein vegt	<i>Xenopus laevis</i>	a (R)	h (A)	f (F)	f (F)		
Individual proteins								
TBX4_MOUSE	TBX-4 (lung, heart)	<i>Mus musculus</i>	a (R)	c (S)	f (F)	f (F)		
TBX5_HUMAN	TBX-5 (heart, limb)	<i>Homo sapiens</i>	a (R)	c (S)	f (F)	f (F)		
OMB_DROME	Optomotor-blind protein	<i>Drosophila melanogaster</i>	a (R)	h (A)	f (F)	h (L)		
TBX5_MOUSE	TBX-5 (heart, limb)	<i>Mus musculus</i>	h (G)	c (S)	f (F)	f (F)		
Conservation score			95.4	91.1	100	92.6		

Table A1.14. Rel homology region family (Highly specific)

The Rel homology region (RHR) is found in the N-terminus of proteins that act at the κ B target site and regulates genes that are commonly involved in cellular defence and differentiation mechanisms³⁴. Binding is through a network of peptide loops that are inserted in the DNA major groove. The C-terminal domains located outside the RHR are highly variable and are likely to function as transcription activators.

Four positions bind the target sequence. Arginine or lysine is found at positions 54 (-.G...-), 56 (-G....-) and 241 (-..Pu-), which all hydrogen bond with guanine. From the variation in target sequence, position 241 may also bind adenine, although guanine is probably favoured because of the possibility of producing a bidentate interaction. Glutamate at position 60 recognises two cytosines in a complex interaction (-C'C'...-). All contacts to the target sequence follow the generic interaction pattern.

With the exception of RELB_HUMAN, which is mutated at position 54, all target-contacting positions are very well conserved and experimental analysis shows that substitutions of the target-contacting positions abolish specific binding³⁵. However, because most DNA-binding is through a flexible loop region, the protein may recognise variants of the target sequence by adjusting the loop structure. No disease-related mutations have been reported so far.

Table A1.14

Rel homology region family (Highly specific)

Representative structure: 1nfk chain A

Target-contacting positions: 54* (-G...-)
56* (-G...-)
60* (-C'C'...-)
241* (-Pu...-)

Target sequence: -GGPuPun-

SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1nfk chain A)														
			54*	56*	57	59	60*	63	64	65	77	136	241*	272			
Subfamily 1																	
REL_AVIRE	Rel transformer (p58)	<i>Avian retic. virus</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (K)	c (N)	a (R)	a (K)			
DIF_DROME	Dorsal-related immunity factor (Dif)	<i>Drosophila melanogaster</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	f (T)	g (A)	g (G)	a (K)	a (K)	a (K)	a (K)			
DORS_DROME	Embryonic polarity dorsal protein	<i>Drosophila melanogaster</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (K)	c (N)	a (K)	a (K)			
RELB_CHICK	RelB homolog (frag)	<i>Gallus gallus</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (K)	c (N)	a (K)	a (K)			
KBZF2_CHICK	Nuclear factor xB (p100)	<i>Gallus gallus</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	a (H)	g (G)	a (K)	c (N)	a (K)	a (K)			
TF65_CHICK	Nuclear factor xB (p65)	<i>Gallus gallus</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (R)	c (N)	a (R)	a (K)			
1svcP	Nuclear factor xB (p50)	<i>Homo sapiens</i>	a (R)	a (R)	g (Y)	g (A)	d (E)	b (S)	a (H)	g (G)	a (K)	c (N)	a (K)	a (K)			
REL_HUMAN	C-rel proto-oncogene	<i>Homo sapiens</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (R)	c (N)	a (R)	a (K)			
KBZF2_HUMAN	Nuclear factor xB (p100)	<i>Homo sapiens</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	a (H)	g (G)	a (K)	c (N)	a (K)	a (K)			
TF65_HUMAN	Nuclear factor xB (p65)	<i>Homo sapiens</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (K)	c (N)	a (R)	a (K)			
1nfkA	Nuclear factor xB (p50)	<i>Mus musculus</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	a (H)	g (G)	a (K)	c (N)	a (K)	a (K)			
RELB_XENLA	RelB homolog	<i>Xenopus laevis</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	f (T)	g (G)	a (K)	c (N)	a (K)	a (K)			
TF65_XENLA	RelB homolog	<i>Xenopus laevis</i>	a (R)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (K)	c (N)	a (R)	a (K)			
Individual proteins																	
RELB_HUMAN	Transcription factor RelB (l-rel)	<i>Homo sapiens</i>	e (P)	a (R)	g (Y)	h (C)	d (E)	b (S)	g (A)	g (G)	a (K)	c (N)	a (K)	a (K)			
Conservation score			92.8	100	100	91.4	100	93.6	71.4	100	90.8	93.6	82.7	100			

Table A1.15. Homeodomain family (Multi-specific)

The homeodomain is a conserved structural motif in many eukaryotic proteins that regulate cell development. DNA binding is through use of a helix-turn-helix motif.

Five positions contact the target sequence. The nature of interactions varies between family members. Position 105 resides in the N-terminal tail preceding the HTH motif and binds thymine or adenine from the minor groove (-(T/A)...-) in the Oct-1 (1octC2), Pit-1 (1au7A2), paired (1fjl) and engrailed (1hdd) homeodomains. The position mostly comprises arginine or lysine, except in subfamily 17-18 and we expect the interaction to be conserved in most proteins.

Position 147: valine or isoleucine interacts with thymine (-...T-) in Oct-1, Pit-1, paired and engrailed homeodomains, valine contacts adenine (-..A'.-) in Mat α 1 (1yrnA) and asparagine contacts cytosine (-.C'.-) in Mat α -2 (1yrnB, 1aplA). The position typically consists of hydrophobic residues, or in subfamilies 4 and 17, asparagine.

Position 150: water-mediated interactions are made in the paired and engrailed homeodomains (-...T-) or Mat α -2 homeodomain (-T...-). No direct contacts with bases are observed. The residue position is variable and substitutions between most amino acid types are observed.

Position 151: asparagine interacts with adenine in all homeodomain proteins (-.A..- or -..A/A'.-). The amino acid is very well conserved except in subfamilies 3 and 7 that contain protein fragments. This is the only generic amino acid-base interaction.

Position 154: interactions are only observed in Mat α -2 where arginine contacts guanine (-.G.-), Mat α 1 where methionine contacts cytosine (-C...-) and Pit-1 homeodomain where glutamine contacts adenine (-...A'.-).

Table A1.15

Homeodomain family (Multi-specific)										
Representative structure:	1yrn chain A									
Helix-turn-helix motif:	Residues 128-161									
Target-contacting positions:	105 (-(T/A)...-)	Oct-1/Pit-1 Homeodomain, paired, engrailed								
	147 (-...T-)	Oct-1/Pit-1 Homeodomain, paired, engrailed								
	147 (-.C...-)	Mat α -2								
	147 (-.A'...-)	Mat a1								
	150 (-...T-)	paired, engrailed								
	150 (-T...-)	Mat a-2								
	151* (-.(A/A')...-)	all								
	154 (-.G...-)	Mat a-2								
	154 (-...A'...-)	Pit-1 homeodomain								
	154 (-C...-)	Mat a1								
Target sequence:	-TAAT-	(Paired/engrailed)								
	-AAAT-, -TAAT	(Oct-1 homeodomain)								
	-ACAT-	(Pit-1 homeodomain)								
	-TGTA-	(Mat α -2)								
	-CATC-	(Mat a1)								
	-ATGC-	(Oct-1 POU domain)								
	-ATAC-	(Pit-1 POU domain)								
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1yrn chain A)							
			104	105	107	147	150	151*	154	155
Subfamily 1										
HM32_CAEEL	Homeobox protein Ceh-32	<i>Caenorhabditis elegans</i>	g (T)	- (-)	l (C)	d (N)	a (K)	d (N)	d (Q)	a (R)
HM33_CAEEL	Homeobox protein Ceh-33	<i>Caenorhabditis elegans</i>	c (S)	- (-)	l (C)	d (N)	a (K)	d (N)	d (Q)	a (R)
HM34_CAEEL	Homeobox protein Ceh-34	<i>Caenorhabditis elegans</i>	d (N)	- (-)	l (C)	d (N)	a (K)	d (N)	d (Q)	a (R)
SIX1_HUMAN	Homeobox protein Six1	<i>Homo sapiens</i>	c (S)	- (-)	l (C)	d (N)	a (K)	d (N)	d (Q)	a (R)
SIX3_MOUSE	Homeobox protein Six3	<i>Mus musculus</i>	g (T)	- (-)	l (C)	d (N)	a (K)	d (N)	d (Q)	a (R)
ARE3_MOUSE	Skeletal muscle-specific are binding p	<i>Mus musculus</i>	h (V)	- (-)	l (C)	d (N)	a (K)	d (N)	d (Q)	a (R)
Subfamily 2										
BR31_BRARE	Brain-specific homeobox/pou domain	<i>Brachydanio rerio</i>	a (K)	a (R)	c (S)	- (-)	- (-)	- (-)	- (-)	- (-)
OCT1_RAT	Octamer-binding transcription factor 1	<i>Rattus norvegicus</i>	a (K)	a (R)	c (S)	- (-)	- (-)	- (-)	- (-)	- (-)
HM16_XENLA	Homeotic protein Nrl-16 (Nrl-21) (fragr	<i>Xenopus laevis</i>	a (K)	a (R)	c (S)	- (-)	- (-)	- (-)	- (-)	- (-)
HM19_XENLA	Homeotic protein Nrl-19 (fragment)	<i>Xenopus laevis</i>	a (K)	a (R)	c (S)	- (-)	- (-)	- (-)	- (-)	- (-)
Subfamily 3										
KNA1_ARATH	Knotted-like homeobox protein 1	<i>Arabidopsis thaliana</i>	a (K)	a (K)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
KNA2_ARATH	Knotted-like homeobox protein 2	<i>Arabidopsis thaliana</i>	a (K)	a (K)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
KNA3_ARATH	Knotted-like homeobox protein 3	<i>Arabidopsis thaliana</i>	a (R)	a (R)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
STM_ARATH	Homeobox protein shootmeristemless	<i>Arabidopsis thaliana</i>	a (K)	a (K)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
HD1_BRANA	Homeobox protein Hd1	<i>Brassica napus</i>	a (R)	a (R)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
HMB1_SOYBN	Homeobox protein Sbh1	<i>Glycine max</i>	a (K)	a (K)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
MEI1_HUMAN	Homeobox protein Meis1	<i>Homo sapiens</i>	a (K)	a (R)	h (I)	d (N)	h (I)	d (N)	a (R)	a (R)
MEI3_HUMAN	Homeobox protein Meis3	<i>Homo sapiens</i>	a (K)	a (R)	h (I)	d (N)	h (I)	d (N)	a (R)	a (R)
HKN1_LYCES	Homeotic protein knotted-1 (Tkn1)	<i>Lycopersicon esculentum</i>	a (K)	a (K)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
HKNL_MALDO	Homeobox protein knotted-1-like	<i>Malus domestica</i>	a (R)	a (R)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
MEI3_MOUSE	Homeobox protein Meis3	<i>Mus musculus</i>	a (K)	a (R)	h (I)	d (N)	h (I)	d (N)	a (R)	a (R)
TGIF_MOUSE	5'-tg-3' interacting factor	<i>Mus musculus</i>	a (R)	a (R)	d (N)	d (N)	h (I)	d (N)	a (R)	a (R)
CUP9_YEAST	Homeobox protein Cup9	<i>Saccharomyces cerevisiae</i>	a (R)	a (R)	d (N)	d (N)	h (I)	d (N)	a (R)	a (R)
YGJ6_YEAST	Homeobox protein in Prp20	<i>Saccharomyces cerevisiae</i>	a (K)	a (R)	d (N)	d (N)	h (I)	d (N)	a (R)	a (R)
HKN1_MAIZE	Homeotic protein knotted-1	<i>Zea mays</i>	a (K)	a (K)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
RSH1_MAIZE	Homeobox protein rough sheath 1	<i>Zea mays</i>	a (K)	a (K)	a (K)	d (N)	h (I)	d (N)	a (K)	a (R)
Subfamily 4										
HM20_CAEEL	Homeobox protein Ceh-20	<i>Caenorhabditis elegans</i>	a (K)	a (R)	d (N)	d (N)	h (G)	d (N)	h (I)	a (R)
HM40_CAEEL	Homeobox protein Ceh-40	<i>Caenorhabditis elegans</i>	a (K)	a (R)	d (N)	d (N)	h (G)	d (N)	h (I)	a (R)
EXD_DROME	Homeobox protein extradenticle	<i>Drosophila melanogaster</i>	a (K)	a (R)	d (N)	d (N)	h (G)	d (N)	h (I)	a (R)
Subfamily 5										
HM19_CAEEL	Homeobox protein Ceh-19 (fragment)	<i>Caenorhabditis elegans</i>	f (P)	a (R)	h (A)	g (T)	d (Q)	d (N)	g (T)	a (K)
HM30_CAEEL	Homeobox protein Ceh-30	<i>Caenorhabditis elegans</i>	h (A)	a (R)	h (I)	g (T)	d (Q)	d (N)	g (T)	a (K)
HM1D_DROAN	Homeobox protein Om(1d)	<i>Drosophila ananassae</i>	h (A)	a (R)	h (A)	g (T)	d (Q)	d (N)	g (T)	a (K)
HX11_HUMAN	Homeobox protein Hox-11 (tcl-3 proto	<i>Homo sapiens</i>	f (P)	a (R)	c (S)	g (T)	d (Q)	d (N)	g (T)	a (K)
HX11_MOUSE	Homeobox protein Hox-11 (Tlx-1)	<i>Mus musculus</i>	f (P)	a (R)	c (S)	g (T)	d (Q)	d (N)	g (T)	a (K)
Subfamily 6										
DLX2_ELECQ	Homeobox protein Dlx-2 (fragment)	<i>Eleutherodactylus coqui</i>	f (P)	a (R)	h (I)	h (I)	- (-)	- (-)	- (-)	- (-)
DLX4_ELECQ	Homeobox protein Dlx-4 (fragment)	<i>Eleutherodactylus coqui</i>	f (P)	a (R)	h (I)	h (I)	- (-)	- (-)	- (-)	- (-)
BRN1_RAT	Brain-specific homeobox/pou domain	<i>Rattus norvegicus</i>	a (K)	a (R)	c (S)	h (V)	- (-)	- (-)	- (-)	- (-)
Subfamily 7										
GSC_BRARE	Homeobox protein gooseoid (Zgsc)	<i>Brachydanio rerio</i>	b (H)	a (R)	h (I)	h (V)	a (K)	d (N)	h (A)	a (K)
OTX1_BRARE	Homeobox protein Otx1 (Zotx1)	<i>Brachydanio rerio</i>	e (E)	a (R)	g (T)	h (V)	a (K)	d (N)	h (A)	a (K)
OTX2_BRARE	Homeobox protein Otx2 (Zotx2)	<i>Brachydanio rerio</i>	e (E)	a (R)	g (T)	h (V)	a (K)	d (N)	h (A)	a (K)
UN30_CAEEL	Homeobox protein Unc-30	<i>Caenorhabditis elegans</i>	d (Q)	a (R)	b (H)	h (V)	a (K)	d (N)	h (A)	a (K)
HM36_CAEEL	Homeobox protein Ceh-36	<i>Caenorhabditis elegans</i>	e (E)	a (R)	c (S)	h (V)	a (K)	d (N)	h (A)	a (K)
HM37_CAEEL	Homeobox protein Ceh-37	<i>Caenorhabditis elegans</i>	e (E)	a (R)	g (T)	h (V)	a (K)	d (N)	h (A)	a (K)
GSC_DROME	Homeobox protein gooseoid	<i>Drosophila melanogaster</i>	b (H)	a (R)	h (I)	h (V)	a (K)	d (N)	h (A)	a (K)
HMOC_DROME	Homeotic protein orthodenticle	<i>Drosophila melanogaster</i>	e (E)	a (R)	g (T)	h (V)	a (K)	d (N)	h (A)	a (K)
GSC_CHICK	Homeobox protein gooseoid	<i>Gallus gallus</i>	b (H)	a (R)	h (I)	h (V)	a (K)	d (N)	h (A)	a (K)

CRX_HUMAN	Cone-rod homeobox protein	<i>Homo sapiens</i>	e (E)	a (R)	g (T)	h (V)	a (K)	d (N)	h (A)	a (K)
PIX1_HUMAN	Pituitary homeobox 1	<i>Homo sapiens</i>	d (Q)	a (R)	b (H)	h (V)	a (K)	d (N)	h (A)	a (K)
PIX2_HUMAN	Pituitary homeobox 2	<i>Homo sapiens</i>	d (Q)	a (R)	b (H)	h (V)	a (K)	d (N)	h (A)	a (K)
GSC_MOUSE	Homeobox protein gooseoid	<i>Mus musculus</i>	b (H)	a (R)	h (I)	h (V)	a (K)	d (N)	h (A)	a (K)
PIX1_MOUSE	Pituitary homeobox 1 (P-Otx)	<i>Mus musculus</i>	d (Q)	a (R)	b (H)	h (V)	a (K)	d (N)	h (A)	a (K)
PIX2_MOUSE	Pituitary homeobox 2	<i>Mus musculus</i>	d (Q)	a (R)	b (H)	h (V)	a (K)	d (N)	h (A)	a (K)
PIX3_MOUSE	Pituitary homeobox 3 (homeobox protein)	<i>Mus musculus</i>	d (Q)	a (R)	b (H)	h (V)	a (K)	d (N)	h (A)	a (K)
OTX_STRPU	Homeobox protein Otx (Spotx)	<i>Strongylocentrotus purpuratus</i>	e (E)	a (R)	g (T)	h (V)	a (K)	d (N)	h (A)	a (K)
GSCA_XENLA	Homeobox protein gooseoid isoform	<i>Xenopus laevis</i>	b (H)	a (R)	h (I)	h (V)	a (K)	d (N)	h (A)	a (K)
Subfamily 8										
MAY1_SCHCO	Mating-type protein A-al-Y1	<i>Schizophyllum commune</i>	f (P)	a (R)	a (K)	h (V)	d (Q)	d (N)	a (R)	a (R)
MAY3_SCHCO	Mating-type protein A-al-Y3	<i>Schizophyllum commune</i>	f (P)	a (R)	a (K)	h (V)	d (Q)	d (N)	a (R)	a (R)
Subfamily 9+A185										
DLX3_AMBME	Homeobox protein Dlx-3	<i>Ambystoma mexicanum</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX1_BRARE	Homeobox protein Dlx-1	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX2_BRARE	Homeobox protein Dlx-2	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX3_BRARE	Homeobox protein Dlx-3	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX4_BRARE	Homeobox protein Dlx-4	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX5_BRARE	Homeobox protein Dlx-5	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX6_BRARE	Homeobox protein Dlx-6	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX7_BRARE	Homeobox protein Dlx-7	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX8_BRARE	Homeobox protein Dlx-8	<i>Brachydanio rerio</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
LIM1_BRARE	Homeobox protein Lim-1	<i>Brachydanio rerio</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	c (S)	a (K)
LIM5_BRARE	Homeobox protein Lim-5	<i>Brachydanio rerio</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	c (S)	a (K)
HMDL_BRAFL	Homeobox protein Dll homolog	<i>Branchiostoma floridae</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
HM23_CAEL	Homeobox protein Ceh-23	<i>Caenorhabditis elegans</i>	h (A)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
HMD1_CAEL	Homeobox protein c28a5.4	<i>Caenorhabditis elegans</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
HMDL_DROME	Homeotic distal-less protein (protein b)	<i>Drosophila melanogaster</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX5_CHICK	Homeobox protein Dlx-5	<i>Gallus gallus</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
LIM1_CHICK	Homeobox protein Lim-1	<i>Gallus gallus</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	c (S)	a (K)
DLX2_HUMAN	Homeobox protein Dlx-2	<i>Homo sapiens</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX4_HUMAN	Homeobox protein Dlx-4 (Dlx-4)	<i>Homo sapiens</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX5_HUMAN	Homeobox protein Dlx-5 (fragment)	<i>Homo sapiens</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX1_MOUSE	Homeobox protein Dlx-1	<i>Mus musculus</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX2_MOUSE	Homeobox protein Dlx-2 (Tes-1)	<i>Mus musculus</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX3_MOUSE	Homeobox protein Dlx-3	<i>Mus musculus</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX5_MOUSE	Homeobox protein Dlx-5	<i>Mus musculus</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX7_MOUSE	Homeobox protein Dlx-7	<i>Mus musculus</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
LIM2_MOUSE	Homeobox protein Lim-2	<i>Mus musculus/rattus norvegicus</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	c (S)	a (K)
BOX5_NOTVI	Homeobox protein box-5 (Nvhbox-5)	<i>Notophthalmus viridescens</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
DLX3_NOTVI	Homeobox protein Dlx-3 (box-4) (Nvht)	<i>Notophthalmus viridescens</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
LIM1_XENLA	Homeobox protein Lim-1 (Xlim-1)	<i>Xenopus laevis</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	c (S)	a (K)
HMD1_XENLA	Homeobox protein Dll-1 (Dll) (Xdll)	<i>Xenopus laevis</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
HMD2_XENLA	Homeobox protein Dll-2 (Xdll-2)	<i>Xenopus laevis</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
HMD3_XENLA	Homeobox protein Dll-3 (Xdll-3)	<i>Xenopus laevis</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
HMD4_XENLA	Homeobox protein Dll-4 (Xdll-4)	<i>Xenopus laevis</i>	f (P)	a (R)	h (I)	h (I)	d (Q)	d (N)	c (S)	a (K)
Subfamily 10										
HK51_MOUSE	Homeobox protein Nkx-5.1	<i>Mus musculus</i>	g (T)	a (R)	h (V)	h (I)	d (Q)	d (N)	d (N)	a (K)
YOX1_YEAST	Homeobox protein Yox1	<i>Saccharomyces cerevisiae</i>	a (R)	a (R)	- (-)	h (I)	d (Q)	d (N)	d (Q)	h (A)
Subfamily 11										
HME5_APIME	Homeobox protein H40 (fragment)	<i>Apis mellifera</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	g (T)	a (K)
HM09_CAEL	Homeobox protein Ceh-9 (fragment)	<i>Caenorhabditis elegans</i>	h (A)	a (R)	g (T)	h (I)	d (Q)	d (N)	g (T)	a (K)
HM11_DROME	Homeobox protein Nk-1 (s59/2)	<i>Drosophila melanogaster</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	g (T)	a (K)
EMX1_HUMAN	Homeobox protein Emx1 (fragment)	<i>Homo sapiens</i>	- (-)	a (R)	h (A)	h (V)	d (Q)	d (N)	g (T)	a (K)
EMX2_HUMAN	Homeobox protein Emx2 (fragment)	<i>Homo sapiens</i>	h (I)	a (R)	h (A)	h (V)	d (Q)	d (N)	g (T)	a (K)
SAX1_CHICK	Homeobox protein Sax-1 (Chox-3) (fr)	<i>Gallus gallus</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	g (T)	a (K)
EMX1_MOUSE	Homeobox protein Emx1 (fragment)	<i>Mus musculus</i>	- (-)	a (R)	h (A)	h (V)	d (Q)	d (N)	g (T)	a (K)
SAX1_MOUSE	Homeobox protein Sax-1 (Nkx-1.1)	<i>Mus musculus</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	g (T)	a (K)
Subfamily 12										
HMEN_ANOGA	Segmentation polarity protein engrailed	<i>Anopheles gambiae</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HM17_APIME	Homeobox protein H17 (fragment)	<i>Apis mellifera</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME3_APIME	Homeobox protein E30 (fragment)	<i>Apis mellifera</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME6_APIME	Homeobox protein E60 (fragment)	<i>Apis mellifera</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMEN_ARTSF	Homeobox protein engrailed	<i>Artemia sanfranciscana</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMX1_BOVIN	Homeobox protein Msx-1	<i>Bos taurus</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME1_BRARE	Homeobox protein engrailed-1	<i>Brachydanio rerio</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME2_BRARE	Homeobox protein engrailed-2 (Zf-en-)	<i>Brachydanio rerio</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME3_BRARE	Homeobox protein engrailed-3 (Zf-en-)	<i>Brachydanio rerio</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMM4_BRARE	Homeobox protein Msh-A	<i>Brachydanio rerio</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMM5_BRARE	Homeobox protein Msh-B (fragment)	<i>Brachydanio rerio</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMMC_BRARE	Homeobox protein Msh-C	<i>Brachydanio rerio</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMMD_BRARE	Homeobox protein Msh-D	<i>Brachydanio rerio</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
LIM3_BRARE	Homeobox protein Lim-3	<i>Brachydanio rerio</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
HM07_CAEL	Homeobox protein Ceh-7	<i>Caenorhabditis elegans</i>	a (R)	a (R)	g (T)	h (I)	d (Q)	d (N)	h (I)	a (R)
HM08_CAEL	Homeobox protein Ceh-8	<i>Caenorhabditis elegans</i>	d (N)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
HM10_CAEL	Homeobox protein Ceh-10	<i>Caenorhabditis elegans</i>	b (H)	a (R)	h (I)	h (V)	d (Q)	d (N)	h (A)	a (K)
HM14_CAEL	Homeobox protein Ceh-14	<i>Caenorhabditis elegans</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
HM16_CAEL	Homeobox protein engrailed-like Ceh-	<i>Caenorhabditis elegans</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HM22_CAEL	Homeobox protein Ceh-22	<i>Caenorhabditis elegans</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
PAL1_CAEL	Homeobox protein Pal-1 (Ceh-3)	<i>Caenorhabditis elegans</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
UNC4_CAEL	Homeobox protein Unc-4	<i>Caenorhabditis elegans</i>	g (T)	a (R)	d (N)	h (V)	d (Q)	d (N)	h (A)	a (K)

TTF1_CANFA	Thyroid transcription factor 1	<i>Canis familiaris</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
VSX1_CARAU	Homeobox protein Vsx-1	<i>Carassius auratus</i>	b (H)	a (R)	h (V)	h (V)	d (Q)	d (N)	h (A)	a (K)
1f1jA	Homeobox protein Vsx-1	<i>Drosophila melanogaster</i>	c (S)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (R)
1hdhD	Engrailed homeodomain	<i>Drosophila melanogaster</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMBP_DROME	Homeobox protein bagpipe (Nk-3)	<i>Drosophila melanogaster</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HMCA_DROME	Homeotic caudal protein	<i>Drosophila melanogaster</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMN2_DROME	Homeobox protein Vnd	<i>Drosophila melanogaster</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HMRO_DROME	Homeobox protein rough	<i>Drosophila melanogaster</i>	d (Q)	a (R)	g (T)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMSH_DROME	Muscle segmentation homeobox	<i>Drosophila melanogaster</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMT1_DROME	Muscle-specific homeobox protein tinn	<i>Drosophila melanogaster</i>	f (P)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HMRO_DROVI	Homeobox protein rough	<i>Drosophila virilis</i>	d (Q)	a (R)	g (T)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMH1_DUGTI	Homeobox protein Dth-1	<i>Dugesia tigrina</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HMH2_DUGTI	Homeobox protein Dth-2	<i>Dugesia tigrina</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HBX3_ECHGR	Homeobox protein Eghbx3 (fragment)	<i>Echinococcus granulosus</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
ANF1_CHICK	Homeobox protein Anf-1 (ganf) (fragm	<i>Gallus gallus</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HM7X_CHICK	Homeobox protein Chox-7 (fragment)	<i>Gallus gallus</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMD1_CHICK	Homeobox protein Chox-cad	<i>Gallus gallus</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME1_CHICK	Homeobox protein engrailed-1 (Gg-en	<i>Gallus gallus</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME2_CHICK	Homeobox protein engrailed-2 (Gg-en	<i>Gallus gallus</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMX1_CHICK	Homeobox protein Msx-1 (Chox-7)	<i>Gallus gallus</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMX2_CHICK	Homeobox protein Msx-2 (Chox-8) (G)	<i>Gallus gallus</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMXX_CHICK	Homeobox protein Ghox-7 (Chox-7) (t	<i>Gallus gallus</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HPR1_CHICK	Homeobox protein Prx-1	<i>Gallus gallus</i>	d (N)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
HXDD_CHICK	Homeobox protein Hox-d13 (Chox-4.8	<i>Gallus gallus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (V)	a (K)
LMX1_CHICK	Homeobox protein Lmx-1	<i>Gallus gallus</i>	f (P)	a (R)	h (I)	h (V)	d (Q)	d (N)	h (A)	a (K)
LIM_HALRO	Homeobox protein Lim (hrLim)	<i>Halocynthia roretzi</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
HMEN_HELTR	Homeobox protein Ht-en (fragment)	<i>Helobdella triserialis</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
LX10_HELTR	Homeobox protein Lox10 (fragment)	<i>Helobdella triserialis</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
CDX1_HUMAN	Homeobox protein Cdx-1	<i>Homo sapiens</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
CDX2_HUMAN	Homeobox protein Cdx-2	<i>Homo sapiens</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
CDX4_HUMAN	Homeobox protein Cdx-4	<i>Homo sapiens</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
CRT1_HUMAN	Cartilage homeoprotein 1 (Cart-1)	<i>Homo sapiens</i>	b (H)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
GBX2_HUMAN	Homeobox protein Gbx-2 (fragment)	<i>Homo sapiens</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HK25_HUMAN	Homeobox protein Nkx-2.5	<i>Homo sapiens</i>	f (P)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HK31_HUMAN	Homeobox protein Nkx-3.1	<i>Homo sapiens</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HME1_HUMAN	Homeobox protein engrailed-1 (Hu-en)	<i>Homo sapiens</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME2_HUMAN	Homeobox protein engrailed-2 (Hu-en)	<i>Homo sapiens</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMX1_HUMAN	Homeobox protein Msx-1 (Hox-7)	<i>Homo sapiens</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMX2_HUMAN	Homeobox protein Msx-2 (Hox-8)	<i>Homo sapiens</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HXBD_HUMAN	Homeobox protein Hox-b13	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (V)	a (K)
HXDD_HUMAN	Homeobox protein Hox-d13 (Hox-4i)	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (V)	a (K)
LH2_HUMAN	Homeobox protein Lh-2	<i>Homo sapiens</i>	l (M)	a (R)	c (S)	h (V)	d (Q)	d (N)	h (A)	a (K)
HK22_MESAU	Homeobox protein Bkx-2.2	<i>Mesocricetus auratus</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
LM12_MESAU	Homeobox protein Lmx-1.2	<i>Mesocricetus auratus</i>	f (P)	a (R)	h (I)	h (V)	d (Q)	d (N)	h (A)	a (K)
LMX1_MESAU	Homeobox protein Lmx-1	<i>Mesocricetus auratus</i>	a (R)	a (R)	h (I)	h (V)	d (Q)	d (N)	h (A)	a (K)
CDX1_MOUSE	Homeobox protein Cdx-1	<i>Mus musculus</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
CDX4_MOUSE	Homeobox protein Cdx-4	<i>Mus musculus</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
CX10_MOUSE	Homeobox protein Chx10	<i>Mus musculus</i>	b (H)	a (R)	h (I)	h (V)	d (Q)	d (N)	h (A)	a (K)
GBX2_MOUSE	Homeobox protein Gbx-2 (Stra7)	<i>Mus musculus</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
GSH2_MOUSE	Homeobox protein Gsh-2	<i>Mus musculus</i>	l (M)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (V)	a (K)
GSH1_MOUSE	Homeobox protein Gsh-1	<i>Mus musculus</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (V)	a (K)
HEX1_MOUSE	Anterior-restricted homeobox protein	<i>Mus musculus</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HK23_MOUSE	Homeobox protein Nkx-2.3	<i>Mus musculus</i>	f (P)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HK25_MOUSE	Homeobox protein Nkx-2.5	<i>Mus musculus</i>	f (P)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HK31_MOUSE	Homeobox protein Nkx-3.1	<i>Mus musculus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HK32_MOUSE	Homeobox protein Nkx-3.2	<i>Mus musculus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HME1_MOUSE	Homeobox protein engrailed-1 (Mo-en)	<i>Mus musculus</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HME2_MOUSE	Homeobox protein engrailed-2 (Mo-en)	<i>Mus musculus</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HXBD_MOUSE	Homeobox protein Hox-b13	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (V)	a (K)
LIM4_MOUSE	Lim homeobox protein 4 (fragment)	<i>Mus musculus</i>	f (P)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
MSX3_MOUSE	Homeobox protein Msx-3	<i>Mus musculus</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
CDX1_RAT	Homeobox protein Cdx-1	<i>Rattus norvegicus</i>	h (Y)	a (R)	h (V)	h (I)	d (Q)	d (N)	h (A)	a (K)
CRT1_RAT	Cartilage homeoprotein 1 (Cart-1)	<i>Rattus norvegicus</i>	b (H)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
LH2_RAT	Homeobox protein Lh-2	<i>Rattus norvegicus</i>	f (P)	a (R)	c (S)	h (V)	d (Q)	d (N)	h (A)	a (K)
MSX2_RAT	Homeobox protein Msx-2 (Hox-8.1) (fr	<i>Rattus norvegicus</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMEN_SCHAM	Homeobox protein engrailed (G-en) (fr	<i>Schistocerca americana</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMEN_TRIGR	Homeobox protein engrailed (Su-hb-e)	<i>Tripneustes gratilla</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
ANF1_XENLA	Homeobox protein Anf-1 (Xanf-1)	<i>Xenopus laevis</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
ANF2_XENLA	Homeobox protein Anf-2 (Xanf-2)	<i>Xenopus laevis</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
CRT1_XENLA	Cartilage homeoprotein 1 (Cart-1) (Xc	<i>Xenopus laevis</i>	b (H)	a (R)	g (T)	h (V)	d (Q)	d (N)	h (A)	a (K)
HK25_XENLA	Homeobox protein Nkx-2.5	<i>Xenopus laevis</i>	f (P)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HMEB_XENLA	Homeobox protein engrailed-1b (En-1)	<i>Xenopus laevis</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMEC_XENLA	Homeobox protein engrailed-2a (En-2)	<i>Xenopus laevis</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HMED_XENLA	Homeobox protein engrailed-2b (En-2)	<i>Xenopus laevis</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	h (A)	a (K)
HNK2_XENLA	Homeobox protein Xenk-2	<i>Xenopus laevis</i>	a (R)	a (R)	h (L)	h (I)	d (Q)	d (N)	h (Y)	a (K)
HX71_XENLA	Homeobox protein Xhox-7.1 (fragment)	<i>Xenopus laevis</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
HX7P_XENLA	Homeobox protein Xhox-7.1' (fragmen	<i>Xenopus laevis</i>	f (P)	a (R)	f (P)	h (I)	d (Q)	d (N)	h (A)	a (K)
MIX1_XENLA	Homeobox protein Mix.1	<i>Xenopus laevis</i>	a (K)	a (R)	f (F)	h (V)	d (Q)	d (N)	h (A)	a (K)
MIX2_XENLA	Homeobox protein Mix.2	<i>Xenopus laevis</i>	a (K)	a (R)	f (F)	h (V)	d (Q)	d (N)	h (A)	a (K)
Subfamily 13										
HXA5_AMBME	Homeobox protein Hox-a5 (fragment)	<i>Ambystoma mexicanum</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	l (M)	a (K)
HXA9_AMBME	Homeobox protein Hox-a9 (fragment)	<i>Ambystoma mexicanum</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	l (M)	a (K)
HXB1_AMBME	Homeobox protein Hox-b1 (AHox1) (fr	<i>Ambystoma mexicanum</i>	h (I)	a (R)	d (N)	h (I)	d (Q)	d (N)	l (M)	a (K)
H114_BRARE	Homeobox protein Hox-114	<i>Brachydanio rerio</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	l (M)	a (K)
HXA1_BRARE	Homeobox protein Hox-a1	<i>Brachydanio rerio</i>	h (I)	a (R)	d (N)	h (I)	d (Q)	d (N)	l (M)	a (K)

HXB5_BRARE	Homeobox protein Hox-b5 (Zf-21)	<i>Brachydanio rerio</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB6_BRARE	Homeobox protein Hox-b6 (Zf-22)	<i>Brachydanio rerio</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC5_BRARE	Homeobox protein Hox-c5 (Hox-3.4) (<i>Brachydanio rerio</i>)	<i>Brachydanio rerio</i>	c (S)	a (R)	c (S)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD4_BRARE	Homeobox protein Hox-d4 (Zf-13) (<i>Brachydanio rerio</i>)	<i>Brachydanio rerio</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
ISL1_BRARE	Insulin gene enhancer protein isl-1 (<i>Brachydanio rerio</i>)	<i>Brachydanio rerio</i>	h (V)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (C)	a (K)
ISL2_BRARE	Insulin gene enhancer protein isl-2 (<i>Brachydanio rerio</i>)	<i>Brachydanio rerio</i>	h (V)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (C)	a (K)
ISL3_BRARE	Insulin gene enhancer protein isl-3 (<i>Brachydanio rerio</i>)	<i>Brachydanio rerio</i>	h (V)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (C)	a (K)
HOX3_BRAFL	Homeobox protein Hox3	<i>Branchiostoma floridae</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
EGL5_CAEEL	Homeobox protein Egl-5	<i>Caenorhabditis elegans</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HM13_CAEEL	Homeobox protein Ceh-13 (fragment)	<i>Caenorhabditis elegans</i>	d (N)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
LI39_CAEEL	Homeobox protein Lin-39	<i>Caenorhabditis elegans</i>	d (Q)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
MAB5_CAEEL	Homeobox protein Mab-5	<i>Caenorhabditis elegans</i>	g (T)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
VAB7_CAEEL	Homeobox protein Vab-7	<i>Caenorhabditis elegans</i>	a (R)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXA9_CAVPO	Homeobox protein Hox-a9 (Hox-1.7) (<i>Cavia porcellus</i>)	<i>Cavia porcellus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA7_COTJA	Homeobox protein Hox-a7 (Quox-1)	<i>Coturnix coturnix japonica</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB1_CYPCA	Homeobox protein Hox-b1	<i>Cyprinus carpio</i>	h (I)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
HMAA_DROME	Homeobox protein abdominal-A	<i>Drosophila melanogaster</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HMAB_DROME	Homeobox protein abdominal-b (P3)	<i>Drosophila melanogaster</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HMAN_DROME	Homeotic antennapedia protein	<i>Drosophila melanogaster</i>	a (R)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HMDF_DROME	Homeotic deformed protein	<i>Drosophila melanogaster</i>	d (Q)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HMH2_DROME	Homeobox protein H2.0	<i>Drosophila melanogaster</i>	c (S)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (M)	a (K)
HMUX_DROME	Homeotic ultrabithorax protein	<i>Drosophila melanogaster</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA2_CHICK	Homeobox protein Hox-a2	<i>Gallus gallus</i>	h (L)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXA4_CHICK	Homeobox protein Hox-a4 (Chox-1.4)	<i>Gallus gallus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA9_CHICK	Homeobox protein Hox-a9 (fragment)	<i>Gallus gallus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXAB_CHICK	Homeobox protein Hox-a11 (Ghox-1i)	<i>Gallus gallus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB1_CHICK	Homeobox protein Hox-b1 (Ghox-lab)	<i>Gallus gallus</i>	h (I)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB3_CHICK	Homeobox protein Hox-b3 (Chox-2.7)	<i>Gallus gallus</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB4_CHICK	Homeobox protein Hox-b4 (Chox-z)	<i>Gallus gallus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD4_CHICK	Homeobox protein Hox-d4 (Chox-a)	<i>Gallus gallus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD8_CHICK	Homeobox protein Hox-d8 (Chox-m)	<i>Gallus gallus</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXDA_CHICK	Homeobox protein Hox-d10 (Chox-4.5)	<i>Gallus gallus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXDB_CHICK	Homeobox protein Hox-d11 (Chox-4.6)	<i>Gallus gallus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXDC_CHICK	Homeobox protein Hox-d12 (Chox-4.7)	<i>Gallus gallus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
ISL1_CHICK	Insulin gene enhancer protein isl-1 (<i>Gallus gallus</i>)	<i>Gallus gallus</i>	h (V)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (C)	a (K)
HOX1_HALRO	Homeobox protein Ahox1	<i>Halocynthia roretzi</i>	d (N)	a (R)	h (V)	h (I)	d (Q)	d (N)	I (M)	a (K)
HML2_HELRO	Homeobox Lox2 protein (fragment)	<i>Helobdella robusta</i>	a (R)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
EVX1_HUMAN	Homeobox even-skipped homolog pro <i>Homo sapiens</i>	<i>Homo sapiens</i>	h (Y)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
EVX2_HUMAN	Homeobox even-skipped homolog pro <i>Homo sapiens</i>	<i>Homo sapiens</i>	h (Y)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HB9_HUMAN	Homeobox protein Hb9	<i>Homo sapiens</i>	f (P)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HLX1_HUMAN	Homeobox protein Hlx1 (homeobox pr <i>Homo sapiens</i>)	<i>Homo sapiens</i>	c (S)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXA1_HUMAN	Homeobox protein Hox-a1 (Hox-1f)	<i>Homo sapiens</i>	h (V)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA4_HUMAN	Homeobox protein Hox-a4 (Hox-1d) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	a (R)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA5_HUMAN	Homeobox protein Hox-a5 (Hox-1c)	<i>Homo sapiens</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA9_HUMAN	Homeobox protein Hox-a9 (Hox-1g) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXAA_HUMAN	Homeobox protein Hox-a10 (Hox-1h) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB1_HUMAN	Homeobox protein Hox-b1 (Hox-2i)	<i>Homo sapiens</i>	h (L)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB2_HUMAN	Homeobox protein Hox-b2 (Hox-2h) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	h (L)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXB3_HUMAN	Homeobox protein Hox-b3 (Hox-2g) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB4_HUMAN	Homeobox protein Hox-b4 (Hox-2f) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB5_HUMAN	Homeobox protein Hox-b5 (Hox-2a) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB6_HUMAN	Homeobox protein Hox-b6 (Hox-2b) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB7_HUMAN	Homeobox protein Hox-b7 (Hox-2c) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB9_HUMAN	Homeobox protein Hox-b9 (Hox-2e) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC4_HUMAN	Homeobox protein Hox-c4 (Hox-3e) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC5_HUMAN	Homeobox protein Hox-c5 (Hox-3d) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	c (S)	a (R)	c (S)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC6_HUMAN	Homeobox protein Hox-c6 (Hox-3c) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	h (G)	a (R)	h (I)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXCB_HUMAN	Homeobox protein Hox-c11	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD3_HUMAN	Homeobox protein Hox-d3 (Hox-4a)	<i>Homo sapiens</i>	h (V)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD4_HUMAN	Homeobox protein Hox-d4 (Hox-4b) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD9_HUMAN	Homeobox protein Hox-d9 (Hox-4c) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXDA_HUMAN	Homeobox protein Hox-d10 (Hox-4d) (<i>Homo sapiens</i>)	<i>Homo sapiens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
IPF1_HUMAN	Insulin promoter factor 1 (Ipf-1)	<i>Homo sapiens</i>	g (T)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
MOX1_HUMAN	Homeobox protein Mox-1	<i>Homo sapiens</i>	e (E)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
MOX2_HUMAN	Homeobox protein Mox-2	<i>Homo sapiens</i>	e (E)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXA4_LINSA	Homeobox protein Hox-a4 (Ishox 4) (<i>Lineus sanguineus</i>)	<i>Lineus sanguineus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
DBX_MOUSE	Homeobox protein Dbx	<i>Mus musculus</i>	a (R)	a (R)	h (V)	h (I)	d (Q)	d (N)	I (M)	a (K)
EVX1_MOUSE	Homeobox even-skipped homolog pro <i>Mus musculus</i>	<i>Mus musculus</i>	h (Y)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
EVX2_MOUSE	Homeobox even-skipped homolog pro <i>Mus musculus</i>	<i>Mus musculus</i>	h (Y)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXA2_MOUSE	Homeobox protein Hox-a2 (Hox-1.11) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (L)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXA3_MOUSE	Homeobox protein Hox-a3 (Hox-1.5) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA4_MOUSE	Homeobox protein Hox-a4 (Hox-1.4) (<i>Mus musculus</i>)	<i>Mus musculus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA6_MOUSE	Homeobox protein Hox-a6 (Hox-1.2) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA7_MOUSE	Homeobox protein Hox-a7 (Hox-1.1) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA9_MOUSE	Homeobox protein Hox-a9 (Hox-1.7) (<i>Mus musculus</i>)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXAA_MOUSE	Homeobox protein Hox-a10 (Hox-1.8) (<i>Mus musculus</i>)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXAB_MOUSE	Homeobox protein Hox-a11 (Hox-1.9) (<i>Mus musculus</i>)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB1_MOUSE	Homeobox protein Hox-b1 (Hox-2.9) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (L)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB3_MOUSE	Homeobox protein Hox-b3 (Hox-2.7) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB4_MOUSE	Homeobox protein Hox-b4 (Hox-2.6) (<i>Mus musculus</i>)	<i>Mus musculus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB8_MOUSE	Homeobox protein Hox-b8 (Hox-2.4) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB9_MOUSE	Homeobox protein Hox-b9 (Hox-2.5) (<i>Mus musculus</i>)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC4_MOUSE	Homeobox protein Hox-c4 (Hox-3.5) (<i>Mus musculus</i>)	<i>Mus musculus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC6_MOUSE	Homeobox protein Hox-c6 (Hox-3.3) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (G)	a (R)	h (I)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC8_MOUSE	Homeobox protein Hox-c8 (Hox-3.1) (<i>Mus musculus</i>)	<i>Mus musculus</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC9_MOUSE	Homeobox protein Hox-c9 (Hox-3.2) (<i>Mus musculus</i>)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)

HXCA_MOUSE	Homeobox protein Hox-c10 (Hox-3.6)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD1_MOUSE	Homeobox protein Hox-d1 (Hox-4.9)	<i>Mus musculus</i>	h (I)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD3_MOUSE	Homeobox protein Hox-d3 (Hox-4.1)	<i>Mus musculus</i>	h (V)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD4_MOUSE	Homeobox protein Hox-d4 (Hox-4.2)	<i>Mus musculus</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB8_MOUSE	Homeobox protein Hox-b8 (Hox-4.3)	<i>Mus musculus</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD9_MOUSE	Homeobox protein Hox-d9 (Hox-4.4)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXDB_MOUSE	Homeobox protein Hox-d11 (Hox-4.6)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXDC_MOUSE	Homeobox protein Hox-d12 (Hox-4.7)	<i>Mus musculus</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
IPF1_MOUSE	Insulin promoter factor 1 (Ipf-1)	<i>Mus musculus</i>	g (T)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
MOX1_MOUSE	Homeobox protein Mox-1	<i>Mus musculus</i>	e (E)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXC6_NOTVI	Homeobox protein Hox-c6 (NvHox-1)	<i>Notophthalmus viridescens</i>	h (G)	a (R)	h (I)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXDB_NOTVI	Homeobox protein Hox-d11 (NvHox-2)	<i>Notophthalmus viridescens</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
IS2A_ONCTS	Insulin gene enhancer protein Isl-2a (i)	<i>Oncorhynchus tshawytscha</i>	h (V)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (C)	a (K)
IS2B_ONCTS	Insulin gene enhancer protein Isl-2b (i)	<i>Oncorhynchus tshawytscha</i>	h (V)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (C)	a (K)
HXC9_SHEEP	Homeobox protein Hox-c9 (fragment)	<i>Ovis aries</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA2_RAT	Homeobox protein Hox-a2 (Hox-1.11)	<i>Rattus norvegicus</i>	h (L)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
IPF1_RAT	Insulin promoter factor 1 (Ipf-1)	<i>Rattus norvegicus</i>	g (T)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
ISL2_RAT	Insulin gene enhancer protein Isl-2 (isl)	<i>Rattus norvegicus</i>	h (V)	a (R)	h (V)	h (V)	d (Q)	d (N)	I (C)	a (K)
HMB1_STRPU	Homeobox protein Hb1 (Sphbox1)	<i>Strongylocentrotus purpuratus</i>	I (C)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HMB3_TRIGR	Homeobox protein Hb3 (fragment)	<i>Tripneustes gratilla</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HMB4_TRIGR	Homeobox protein Hb4 (fragment)	<i>Tripneustes gratilla</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HB7A_XENLA	Homeobox protein Hox-b7 a (Xlhbox-2)	<i>Xenopus laevis</i>	a (R)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HB7B_XENLA	Homeobox protein Hox-b7 b (Xlhbox-2)	<i>Xenopus laevis</i>	a (R)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HM8_XENLA	Homeobox protein 8 (Xlhbox-8)	<i>Xenopus laevis</i>	g (T)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HX3_XENLA	Homeobox protein Xhox-3	<i>Xenopus laevis</i>	h (Y)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
HXA1_XENLA	Homeobox protein Hox-a1 (Hox.lab2)	<i>Xenopus laevis</i>	h (A)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXA7_XENLA	Homeobox protein Hox-a7 (hnhbox-3)	<i>Xenopus laevis</i>	h (G)	a (R)	g (T)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB4_XENLA	Homeobox protein Hox-b4 (Xhox-1a)	<i>Xenopus laevis</i>	c (S)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB5_XENLA	Homeobox protein Hox-b5 (Xlhbox-4)	<i>Xenopus laevis</i>	h (A)	a (R)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXB9_XENLA	Homeobox protein Hox-b9 (Xlhbox-6)	<i>Xenopus laevis</i>	a (K)	a (R)	f (P)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXC5_XENLA	Homeobox protein Hox-c5 (Xlhbox-5)	<i>Xenopus laevis</i>	c (S)	a (R)	c (S)	h (I)	d (Q)	d (N)	I (M)	a (K)
HXD1_XENLA	Homeobox protein Hox-d1 (Hox.lab1)	<i>Xenopus laevis</i>	h (V)	a (R)	d (N)	h (I)	d (Q)	d (N)	I (M)	a (K)
MOX2_XENLA	Homeobox protein Mox-2	<i>Xenopus laevis</i>	e (E)	a (R)	h (A)	h (V)	d (Q)	d (N)	I (M)	a (K)
Subfamily 14										
BR3A_RAT	Brain-specific homeobox/POU domain	<i>Rattus norvegicus</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	- (-)	- (-)
BRN1_PIG	Brain-specific homeobox/POU domain	<i>Sus scrofa</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	- (-)	- (-)
Subfamily 15										
BR11_BRARE	Brain-specific homeobox/POU domain	<i>Brachydanio rerio</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
POU2_BRARE	Pou domain protein 2	<i>Brachydanio rerio</i>	a (R)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
HM06_CAEEL	Homeobox protein Ceh-6	<i>Caenorhabditis elegans</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
HM18_CAEEL	Homeobox protein Ceh-18	<i>Caenorhabditis elegans</i>	a (K)	a (R)	d (N)	h (V)	I (C)	d (N)	d (Q)	a (K)
UN86_CAEEL	Transcription factor Unc-86	<i>Caenorhabditis elegans</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
IPOU_DROME	Inhibitory pou protein (I-POU)	<i>Drosophila melanogaster</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
PDM1_DROME	Nubbin protein (twain protein)	<i>Drosophila melanogaster</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
PDM2_DROME	Didymous protein (pou domain protein)	<i>Drosophila melanogaster</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
POU1_DUGJA	Pou domain protein 1 (DJPOU1)	<i>Dugesia japonica</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
OCT1_CHICK	Octamer-binding transcription factor 1	<i>Gallus gallus</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
1octC2	Oct-1 homeodomain	<i>Homo sapiens</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
BR3A_HUMAN	Brain-specific homeobox/POU domain	<i>Homo sapiens</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
OCT3A_HUMAN	Octamer-binding transcription factor 3	<i>Homo sapiens</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
OCT1_HUMAN	Octamer-binding transcription factor 1	<i>Homo sapiens</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
OCT2_HUMAN	Octamer-binding transcription factor 2	<i>Homo sapiens</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
PO61_HUMAN	Pou domain, class 6, transcription fact	<i>Homo sapiens</i>	a (R)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	g (T)
OCT11_MOUSE	Octamer-binding transcription factor 1	<i>Mus musculus</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
1au7A2	Pit-1 homeodomain	<i>Rattus norvegicus</i>	a (R)	a (R)	g (T)	h (V)	I (C)	d (N)	d (Q)	a (K)
OCT1_XENLA	Octamer-binding transcription factor 1	<i>Xenopus laevis</i>	a (K)	a (R)	c (S)	h (V)	I (C)	d (N)	d (Q)	a (K)
Subfamily 16										
1apC	Mat a1-2	<i>Saccharomyces cerevisiae</i>	a (R)	h (G)	a (R)	d (N)	c (S)	d (N)	a (R)	a (K)
1ymB	Mat a1-2	<i>Saccharomyces cerevisiae</i>	a (R)	h (G)	a (R)	d (N)	c (S)	d (N)	a (R)	a (K)
MAT2_YEAST	Mating-type protein a1-2	<i>Saccharomyces cerevisiae</i>	a (R)	h (G)	a (R)	d (N)	c (S)	d (N)	a (R)	a (K)
Subfamily 17										
HMPH_CHICK	Homeobox protein Prh	<i>Gallus gallus</i>	h (G)	h (G)	a (R)	g (T)	d (Q)	d (N)	h (A)	a (K)
HMPH_HUMAN	Homeobox protein Prh (homeobox prc)	<i>Homo sapiens</i>	h (G)	h (G)	a (R)	g (T)	d (Q)	d (N)	h (A)	a (K)
HMPH_MOUSE	Homeobox protein Prh (Hex)	<i>Mus musculus</i>	h (G)	h (G)	a (R)	g (T)	d (Q)	d (N)	h (A)	a (K)
Individual proteins										
HXC8_RAT	Homeobox protein Hox-c8 (R4) (fragm)	<i>Rattus norvegicus</i>	- (-)	- (-)	- (-)	h (I)	d (Q)	d (N)	I (M)	a (K)
1ymA	Mat A1	<i>Saccharomyces cerevisiae</i>	- (-)	- (-)	- (-)	h (V)	h (I)	d (N)	I (M)	a (R)
1octA1	Oct-1 POU domain	<i>Homo sapiens</i>								
CDP_CANFA	CCAAT displacement protein	<i>Canis familiaris</i>	f (P)	a (R)	h (V)	d (N)	b (H)	d (N)	c (S)	a (R)
PRH_PETCR	Pathogenesis-related homeodomain p	<i>Petroselinum crispum</i>	d (Q)	a (R)	- (-)	d (N)	d (N)	d (N)	I (W)	c (S)
HK52_MOUSE	Homeobox protein Nkx-5.2 (fragment)	<i>Mus musculus</i>	g (T)	a (R)	h (V)	g (T)	d (Q)	d (N)	d (N)	a (K)
HMBC_DROME	Homeotic bicoid protein (Prd-4)	<i>Drosophila melanogaster</i>	g (T)	a (R)	g (T)	h (I)	a (K)	d (N)	a (R)	a (R)
SPM1_RAT	Sperm 1 pou-domain transcription fact	<i>Rattus norvegicus</i>	a (K)	a (R)	a (R)	h (V)	c (S)	d (N)	d (Q)	I (M)
MTA0_YEAST	Mating-type protein a-1	<i>Saccharomyces cerevisiae</i>	h (G)	a (K)	c (S)	h (V)	I (C)	d (N)	h (I)	a (K)
PNX1_HUMAN	Homeobox protein Pknx1	<i>Homo sapiens</i>	a (R)	h (G)	h (V)	d (N)	h (I)	d (N)	a (R)	a (R)
LIM1_HUMAN	Homeobox protein Lim-1	<i>Homo sapiens</i>	f (P)	h (G)	g (T)	h (V)	d (Q)	d (N)	c (S)	a (K)
HXD3_RAT	Homeobox protein Hox-d3 (R6) (fragm)	<i>Rattus norvegicus</i>	h (G)	I (C)	h (A)	h (I)	d (Q)	d (N)	I (M)	a (K)
Conservation score			52.5	90.9	54.3	76.2	74.6	96.3	54	87.9

Table A1.16. LacI family (Multi-specific)

The lactose repressor family (LacI) comprises a set of homologous transcription regulators in numerous bacterial biosynthetic pathways, including purine, pyrimidine, cytosine and sugar synthesis. Lactose, fructose, and raffinose repressors act as tetramers while remaining proteins function as dimers.

Four positions contact the target sequence in the purine repressor structures (1wetA). Positions 16 and 26 from the helix-turn-helix motif bind in the major groove towards the edge of the target site. Position 16 comprises threonine (subfamilies 3-4), serine (subfamilies 1-2), glutamine (individual), aspartate (individual) and hydrophobic amino acids (subfamilies 5-9). Threonine in 1wetA binds thymine (-....T'...-). Position 26 consists of arginine (subfamilies 1 and 3), serine (subfamily 2), proline (subfamilies 5 and 8), amide residues (subfamily 7) and hydrophobic amino acids (subfamilies 4 and 9). Arginine in 1wetA binds guanine (-..G.....-) in a bidentate interaction.

Positions 54 and 55 from the hinge helix bind in the minor groove at the centre of the target site. A hydrophobic residue is invariably found at position 54; here leucines from the two subunits cooperatively intercalate between the central C.G base-step between the two half-sites (-.....CG-). At position 55, arginine, lysine (subfamilies 1-3, 5--6), hydrophobic amino acids (subfamilies 7--9) or glutamine (subfamily 4) is found. Lysine in 1wetA binds the adenine N3 atom (-.....A.-).

Experimental analysis on the purine repressor shows that substitutions at positions 16 and 26 for other amino acid types leads to loss of transcription regulation of the PurF-lacZ gene. Mutations of position 54 for valine or methionine have no effect, but mutations to any other amino acid abolish binding. The intercalation by this residue position is expected in all LacI repressor complexes^{36,37}. Mutation of position 55 from lysine to alanine weakens the preference for adenine in the target sequence (-.....n.-) because the side chain no longer sterically clashes with the exocyclic N2 atom of guanine³⁸. An extensive mutation analysis along the whole length of the LacI repressor protein shows that binding is either altered or abolished by substitutions at the target contacting positions³⁹.

Potentially significant amino acid substitutions are found at positions 16, 26 and 55. Substitutions are found between different proteins originating from the same organism, and therefore belong to different recognition subfamilies. The corresponding bases in the target sequences differ for the different proteins. The extent of mutation varies. For example, the ribose (RBSR_ECOLI) and trehalose (TRER_ECOLI) repressors from *E. coli* differ only at position 26, while comparison with the galactose repressor (GALR_ECOLI) reveals that substitutions are found at positions 16, 26 and 55. Interestingly, the galactose ultra-induction factor (GALS_ECOLI) and the galactose repressor (GALR_ECOLI), two proteins that have opposing effects on the same biochemical process only differ at position 26.

Target-contacting positions in orthologous proteins from different organisms are generally conserved and therefore belong to the same recognition subfamily. For example, ribose repressors in subfamily 1, purine repressors in subfamily 3 and sucrose repressors in subfamily 4. However, mutations are also observed and some

orthologues have been placed in different subfamilies; for example, maltose repressors in individual proteins.

Table A1.A1.16

LacI family (Multi-specific)								
Representative structure:	1wet chain A							
Helix-turn-helix motif:	Residues 4-22							
Hinge motif:	Residues 49-56							
Target-contacting positions:	16* (-...T...-) 26* (-.G.....-) 54 (-.....C-) 55* (-.....A.-)							
Target sequence:	-ACGCAAAC- (PurR) -GTG(G/TAAAnC- (GalR) -GTG(G/TAAAn- (GalS) -TTGTGAGC- (LacI) -GATAAAAC- (MalR) -(A/TCCGAAAC- (RafR) -nCTAAACC- (ScrR)							
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1wet chain A)					
			16*	20	26*	51	54	55*
Subfamily 1								
RBSR_ECOLI	Ribose repressor	<i>Escherichia coli</i>	c (S)	b (H)	a (R)	h (A)	h (L)	a (K)
RBSR_HAEIN	Ribose repressor	<i>Haemophilus influenzae</i>	c (S)	b (H)	a (R)	h (A)	h (L)	a (K)
Subfamily 2								
TRER_ECOLI	Trehalose repressor	<i>Escherichia coli</i>	c (S)	a (R)	c (S)	h (A)	l (M)	a (R)
SCR_R_STAXY	Sucrose regulator	<i>Staphylococcus xylosum</i>	c (S)	a (R)	c (S)	h (A)	h (L)	a (R)
TRER_SALTY	Trehalose repressor	<i>Salmonella typhimurium</i>	c (S)	a (R)	c (S)	h (A)	l (M)	a (R)
Subfamily 3								
1wetA	Purine repressor	<i>Escherichia coli</i>	g (T)	b (H)	a (R)	h (A)	h (L)	a (K)
PURR_HAEIN	Purine repressor	<i>Haemophilus influenzae</i>	g (T)	b (H)	a (R)	h (A)	h (L)	a (K)
Subfamily 4								
SCR_R_STRMU	Sucrose regulator	<i>Streptococcus mutans</i>	g (T)	a (R)	h (G)	h (A)	h (L)	d (Q)
SACR_LACLA	Sucrose regulator	<i>Lactococcus lactis</i>	g (T)	a (R)	h (G)	h (A)	h (L)	d (Q)
SCR_R_PEDPE	Sucrose regulator	<i>Pediococcus pentosaceus</i>	g (T)	a (R)	h (G)	h (A)	h (L)	d (Q)
Subfamily 5								
MALR_STRCO	Maltose repressor	<i>Streptomyces coelicolor</i>	h (A)	a (R)	f (P)	- (-)	h (L)	a (R)
REGL_STRLI	RegL regulator	<i>Streptomyces lividans</i>	h (A)	a (R)	f (P)	- (-)	h (L)	a (R)
Subfamily 6								
CYTR_ECOLI	Cytosine repressor	<i>Escherichia coli</i>	h (A)	a (R)	e (D)	h (G)	h (V)	a (K)
REGA_CLOAB	RegA regulator	<i>C. acetobutylicum</i>	h (A)	a (R)	e (D)	h (A)	h (L)	a (K)
Subfamily 7								
CCPA_STAXY	Catabolite control	<i>Staphylococcus xylosum</i>	h (A)	a (R)	d (Q)	h (A)	h (L)	h (A)
PEPR_LACDL	PEPR1 regulator	<i>Lactobacillus delbrueckii</i>	h (A)	a (R)	d (N)	h (A)	h (L)	h (A)
Subfamily 8								
GALR_ECOLI	Galactose repressor	<i>Escherichia coli</i>	h (A)	a (R)	f (P)	h (A)	h (L)	h (A)
CCPA_BACSU	Glucose-resistance amylase regulator	<i>Bacillus subtilis</i>	h (A)	a (R)	f (P)	h (A)	h (L)	h (A)
CCPA_BACME	Glucose-resistance amylase regulator	<i>Bacillus megaterium</i>	h (A)	a (R)	f (P)	h (A)	h (L)	h (A)
Subfamily 9								
RBSR_BACSU	Ribose repressor	<i>Bacillus subtilis</i>	h (A)	a (R)	h (G)	h (A)	h (L)	h (Y)
GALR_HAEIN	Galactose repressor	<i>Haemophilus influenzae</i>	h (A)	a (R)	h (L)	h (A)	h (L)	h (A)
Individual proteins								
MALR_STRPN	Maltose repressor	<i>Streptococcus pneumoniae</i>	c (S)	a (R)	c (S)	h (L)	l (W)	h (L)
MALR_CLOBU	Maltose repressor	<i>Clostridium butyricum</i>	c (S)	a (R)	f (P)	h (A)	h (L)	a (K)
MALR_STAXY	Maltose repressor	<i>Staphylococcus xylosum</i>	c (S)	a (R)	f (P)	h (A)	h (L)	h (A)
MALI_ECOLI	Maltose regulator	<i>Escherichia coli</i>	c (S)	h (L)	h (G)	h (A)	h (L)	a (R)
LACI_ECOLI	Lactose repressor	<i>Escherichia coli</i>	d (Q)	a (R)	c (S)	h (A)	h (L)	h (A)
LACI_KLEPN	Lactose repressor	<i>Klebsiella pneumoniae</i>	e (D)	a (R)	e (E)	h (A)	h (L)	h (A)
RAFR_ECOLI	Raffinose repressor	<i>Escherichia coli</i>	g (T)	a (R)	c (S)	h (A)	h (L)	a (K)
DEGA_BACSU	Degradation activator	<i>Bacillus subtilis</i>	g (T)	a (R)	h (G)	h (A)	h (L)	g (T)
CCPA_STRMU	Catabolite control protein A	<i>Streptococcus mutans</i>	h (A)	a (R)	a (K)	h (A)	h (L)	h (A)
GALS_ECOLI	Mgl repressor/galactose ultrainductor	<i>Escherichia coli</i>	h (A)	a (R)	g (T)	h (A)	h (L)	h (A)
ASCG_ECOLI	Cryptic asc repressor	<i>Escherichia coli</i>	h (A)	a (R)	h (G)	h (A)	h (L)	c (S)
GNT_R_ECOLI	Gluconate utilization system GNT-I	<i>Escherichia coli</i>	l (M)	a (R)	e (E)	f (P)	h (L)	c (S)
CSCR_ECOLI	Sucrose (csc) repressor	<i>Escherichia coli</i>	l (M)	a (R)	e (E)	h (A)	l (M)	a (R)
Conservation score			64.9	87.9	54.4	80.3	89.4	57.8

Table A1.17. CAP family (Multi-specific)

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: the catabolite gene regulators (CAP) bind cAMP to control transcription of catabolite sensitive operons and the fumarate/nitrate regulators control genes that are linked to anaerobic electron transport systems. Only the CAP complex structure is available and its target sequence is provided.

Three positions interact with the target site in the CAP complex (subfamily 1). Positions 180 and 185 comprise arginines that bind guanines in bidentate interactions (-G.....- and -...G....- respectively). Position 181 consists of glutamate that interacts with cytosine (-...C'....-). Not all the bases in the target site are contacted and indirect recognition of the sequence is likely as the bound DNA structure is significantly deformed in the CAP complex ⁴⁰.

The three positions are conserved for CAP and mutagenesis studies show that amino acid substitutions alter the specificity of the protein ^{41,42}. Glutamate at position 181 is especially important and is conserved anaerobic regulators in subfamily 2. Mutation of the position to glutamine in CAP is lethal for *E. coli* ⁴³, but surprisingly, substitutions by aromatic amino acids maintains equal specificity as alternative hydrogen bonds are formed ⁴⁴.

CAP (subfamily 1) and anaerobic regulators (subfamilies 2--3) belong to separate recognition subfamilies. Out of the three target-contacting positions, subfamilies 1 and 2 only differ at position 180, where arginine in CAP is replaced by a hydrophobic residue. *E. coli*, *H. influenzae*, and *P. aeruginosa* present both protein types in the current dataset and mutagenesis studies of *E. coli* promoters succeeds in converting a CAP binding site to a FNR site with a single base substitution in the half-site (-X.....-) ⁴⁵. The anaerobic regulators in subfamily 3 differ at position 181 where, valine is found in place of glutamate. The subfamily members originate from organisms presenting single proteins only and therefore the significance of the mutation with respect to CAP is unknown, however, a similar change in the binding site is expected.

Table A1.17

CAP family (Multi-specific)							
Representative structure:	1ber chain A						
Helix-turn-helix motif:	Residues 169-192						
Target-contacting positions:	180* (-G.....-)						
	181 (-...C'.....-)						
	185* (-...G.....-)						
Target sequence:	-TGTGA...- (CAP)						
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1ber chain A)				
			179	180*	181	182	185*
Subfamily 1							
1berA	Catabolite gene activator	<i>Escherichia coli</i>	c (S)	a (R)	e (E)	g (T)	a (R)
2cgpC	Catabolite gene activator	<i>Escherichia coli</i>	c (S)	a (R)	e (E)	g (T)	a (R)
CRP_HAEIN	Catabolite gene activator	<i>Haemophilus influenzae</i>	c (S)	a (R)	e (E)	g (T)	a (R)
VFR_PSEAE	Catabolite gene activator	<i>Pseudomonas aeruginosa</i>	c (S)	a (R)	e (E)	i (M)	a (R)
Subfamily 2							
HLYX_ACTPL	Regulator HLYX	<i>Actinobacillus pleuropneumoniae</i>	g (T)	h (I)	e (E)	g (T)	a (R)
FIXK_AZOCA	Nitrogen fixation regulator FIXK	<i>Azorhizobium caulinodans</i>	g (T)	h (I)	e (E)	g (T)	a (R)
FIXK_BRAJA	Nitrogen fixation regulator FIXK	<i>Bradyrhizobium japonicum</i>	g (T)	h (I)	e (E)	g (T)	a (R)
FNR_ECOLI	Fumarate/nitrate reduction regulator	<i>Escherichia coli</i>	g (T)	h (V)	e (E)	g (T)	a (R)
FNR_HAEIN	Fumarate/nitrate reduction regulator	<i>Haemophilus influenzae</i>	g (T)	h (V)	e (E)	g (T)	a (R)
ANR_PSEAE	Transcriptional activator protein ANR	<i>Pseudomonas aeruginosa</i>	h (A)	h (V)	e (E)	g (T)	a (R)
FNRA_PSEST	Fumarate/nitrate reduction regulator	<i>Pseudomonas stutzeri</i>	h (A)	h (V)	e (E)	g (T)	a (R)
FNRL_RHOSH	Fumarate/nitrate reduction regulator	<i>Rhodobacter sphaeroides</i>	g (T)	h (L)	e (E)	g (T)	a (R)
FIXK_RHIME	Nitrogen fixation regulator FIXK	<i>Rhizobium meliloti</i>	g (T)	h (I)	e (E)	g (T)	a (R)
AADR_RHOPA	Anaerobic aromatic degradation reg.	<i>Rhodopseudomonas palustris</i>	g (T)	h (I)	e (E)	g (T)	a (R)
ETRA_SHEPU	Electron transport regulator A	<i>Shewanella putrefaciens</i>	g (T)	h (V)	e (E)	g (T)	a (R)
Subfamily 3							
NTCA_ANASP	Global nitrogen regulator	<i>Anabaena sp</i>	g (T)	a (R)	h (V)	g (T)	a (R)
NTCA_SYNP7	Global nitrogen regulator	<i>Synechococcus sp</i>	g (T)	a (R)	h (V)	g (T)	a (R)
NTCA_SYNY3	Global nitrogen regulator	<i>Synechocystis sp</i>	g (T)	a (R)	h (V)	g (T)	a (R)
Individual prot							
FNR_BACSU	Anaerobic regulator	<i>Bacillus subtilis</i>	h (A)	a (R)	e (E)	c (S)	a (R)
FLP_LACCA	Fumarate/nitrate reduction regulator	<i>Lactobacillus casei</i>	g (T)	f (P)	e (E)	g (T)	a (R)
CLP_XANCP	Catabolite gene activator-like	<i>Xanthomonas campestris</i>	i (C)	h (A)	d (Q)	i (M)	a (R)
Conservation score			70.4	59.3	80.1	83.4	100

Table A1.18. $\gamma\delta$ -resolvase family (Multi-specific)

The $\gamma\delta$ -resolvase family contains specific recombinases that convert negatively supercoiled circular DNA with two directly repeated recombination sites into two interlinked circular DNA rings. Binding is mainly through a helix-turn-helix motif at the outer edge of the target site and a hinge helix structure that binds in the minor groove at the centre.

Five positions interact with the target in the structure 1gdt. Position 130 consists of arginine (subfamily 1), which binds the N3 atom of stacked adenines (-.....A'A'..-) in the minor groove, and hydrophobic amino acids (subfamilies 2--4). Position 142 is occupied mainly by basic residues (subfamilies 1--4) that bind minor groove acceptor atoms of thymine and adenine in a complex interaction (-.....T'A'.....-). Serine, glutamate, proline and alanine are also found at the position (individual proteins). Position 172 is dominated by arginine (subfamilies 1--3) which binds guanine in a bidentate interaction (-.G.....-.) and hydrophobic amino acids (subfamily 4). Position 173 has serine (subfamilies 1 and 4) that binds guanine (-.... G'.....-) in a bifurcated bond, glutamine (subfamily 2), hydrophobic residues (subfamily 3) and others. The last position, 176 contains a well-conserved tyrosine that uses the phenol ring to contact the methyl group. The DNA is substantially bent in the structure 1gdtA and therefore the flexibility of the nucleic acid is likely play an important role in indirect sequence recognition.

All five positions are mutated in different resolvases from the same organism (*eg* RES4_ECOLI and TNP2_ECOLI) and positions are conserved for orthologues across species (*eg* RES4_ECOLI and TNPT_PSEPU). Experimental analysis shows that in $\gamma\delta$ -resolvase, mutation of amino acid positions 130, 142, 172 and 176 decrease binding affinity for the *res* site and alter base specificity^{46,47}.

Table A1.18

$\gamma\delta$-resolvase family (Multi-specific)											
Representative structure:		1gdt chain A									
Helix-turn-helix motif:		Residues 161-180									
Hinge helix:		Residues 102-136									
Target-contacting positions:		130 (-.....A'A'..-) 142* (-.....T(A/T').....-) 172* (-..G.....-) 173* (-...G'.....-) 176 (-..T.....-)									
Target sequence:		-(C/T)GTCCGA(A/T)A(T/A)(A/T)Tnn- (Res recombination site)									
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1gdt chain A)								
			122	130	141	142*	171	172*	173*	176	177
Subfamily 1											
1gdtA	$\gamma\delta$ -resolvase	<i>Escherichia coli</i>	h (I)	a (R)	h (G)	a (R)	h (A)	a (R)	c (S)	h (Y)	a (K)
1resA	$\gamma\delta$ -resolvase	<i>Escherichia coli</i>	- (-)	- (-)	- (-)	- (-)	h (A)	a (R)	c (S)	h (Y)	a (K)
RES4_ECOLI	Recombinase	<i>Escherichia coli</i>	h (I)	a (R)	h (G)	a (R)	h (A)	a (R)	c (S)	h (Y)	a (K)
TNP3_ECOLI	Transposon TN3 resolvase	<i>Escherichia coli</i>	h (I)	a (R)	h (G)	a (R)	h (A)	a (R)	c (S)	h (Y)	a (K)
TNPT_PSEPU	Resolvase/recombinase	<i>Pseudomonas putida</i>	h (I)	a (R)	h (G)	a (R)	h (G)	a (R)	c (S)	h (Y)	a (K)
Subfamily 2											
TNP2_ECOLI	Transposon TN21 resolvase	<i>Escherichia coli</i>	h (I)	h (I)	h (G)	a (R)	c (S)	a (R)	e (E)	h (Y)	d (Q)
TNP5_PSEAE	Transposon TN501 resolvase	<i>Pseudomonas aeruginosa</i>	h (I)	h (I)	h (G)	a (R)	c (S)	a (R)	e (E)	h (Y)	d (Q)
Subfamily 3											
DNIV_BPMU	DNA-invertase	<i>Bacteriophage mu</i>	h (I)	h (L)	f (P)	a (K)	h (A)	h (L)	c (S)	h (Y)	a (K)
DNIV_BPP1	DNA-invertase	<i>Bacteriophage p1</i>	h (I)	h (L)	h (G)	a (R)	h (A)	h (V)	c (S)	h (Y)	a (K)
Individual proteins											
DNIV_SALAE	DNA-invertase	<i>Salmonella abortus-equi</i>	h (I)	h (L)	h (G)	b (H)	h (G)	h (V)	c (S)	h (Y)	a (R)
TNP6_ENTFC	Transposon TN1546 resolvase	<i>Enterococcus faecium</i>	h (I)	h (I)	h (Y)	b (H)	c (S)	a (R)	h (A)	h (Y)	a (R)
Y4LS_RHISN	Integrase-like protein (Y4LS)	<i>Rhizobium sp</i>	a (R)	h (I)	h (G)	a (R)	c (S)	a (R)	h (G)	h (Y)	d (Q)
Y4CG_RHISN	Invertase-like protein (Y4CG)	<i>Rhizobium sp</i>	h (I)	- (-)	h (A)	e (E)	a (R)	a (R)	f (P)	h (I)	a (R)
BINL_STAAU	Transposon TN552 resolvase	<i>Staphylococcus aureus</i>	h (I)	a (R)	e (E)	a (K)	c (S)	a (R)	g (T)	h (Y)	a (R)
UVP1_ECOLI	UVP1 protein	<i>Escherichia coli</i>	d (N)	h (L)	h (G)	a (R)	c (S)	a (R)	i (M)	h (Y)	a (R)
PAA4_ECOLI	Resolvase	<i>Escherichia coli</i>	h (A)	h (Y)	h (G)	a (R)	c (S)	f (P)	c (S)	a (K)	a (R)
DNIV_ECOLI	DNA-invertase PIN	<i>Escherichia coli</i>	h (I)	h (L)	h (G)	a (R)	h (G)	h (V)	c (S)	h (Y)	a (K)
TNP7_ENTFA	Transposon TN917 resolvase	<i>Enterococcus faecalis</i>	h (I)	h (L)	f (P)	c (S)	h (L)	a (K)	g (T)	h (Y)	a (R)
RESP_CLOPE	Resolvase (ORF8)	<i>Clostridium perfringens</i>	h (I)	h (L)	f (P)	c (S)	c (S)	a (R)	h (A)	h (Y)	a (R)
BIN3_STAAU	Transposon TN552 resolvase (BIN3)	<i>Staphylococcus aureus</i>	c (S)	h (I)	e (D)	f (P)	g (T)	a (R)	d (Q)	h (Y)	a (R)
TNP0_ECOLI	Transposon TN2501 resolvase	<i>Escherichia coli</i>	h (L)	h (I)	c (S)	a (R)	g (T)	a (R)	d (Q)	h (L)	a (R)
Conservation score			76.3	37.1	45.4	46.3	47.3	55.8	47.1	63.7	57.1

Table A1.19. C₂H₂-zinc finger -family (Multi-specific)

The C₂H₂-zinc finger is one of the most common DNA-binding motifs in eukaryotic transcription regulators. The probe helix commonly recognises a three base-pair DNA sequence using a near one-to-one interactions between amino acids and bases. The following abbreviations are used in the table: ZF - zinc finger, TF - transcription factor.

There are four target-contacting positions. In most structures, position 118 contacts -..n-, either position 120 or 121 contacts -.n.- and position 124 contacts -n..-. The motifs in structure 2drp are exceptions. In 2drpA1, position 120 interacts with -..n- instead. In 2drpA2, position 118 contacts -.n.-, position 120 with -..n- and position 121 with -n..-.

In almost all structures, amino acid-base interactions follow the generic pattern on a one-to-one basis. Therefore, set (a) residues interact with guanine, set (c) with adenine, and occasionally set (d) with cytosine. Sequence recognition is very dependent on hydrogen bond interactions and amino acids from sets (e--h) rarely interact if they are present. In addition serine, commonly found at position 120, does not often interact either ⁴⁸. This is why residue positions often appear to miss some interactions and why the central base-pair from the subsite is either contacted by position 120 or 121 depending on the motif sequence. There are unique exceptions in lubdC3 and lubdC4 where leucine and threonine interact with thymine.

The target-contacting positions mutate between all possible amino acid sets. This results in very diverse combinations of amino acid sequences at these positions and therefore a large number of recognition subfamilies. For the reasons given above, hydrogen-bonding residues are most common. Position 120 is an exception, which has an especially high concentration of serine residues ⁴⁹. *In vivo* and *in vitro* experiments have shown that sequence specificity can be controlled through point mutations are target-contacting positions in Zif268-like proteins ⁵⁰⁻⁵².

Table A1.19

Zinc finger family (Multi-specific)								
Representative structure:	1aay chain A							
Probe α -helix:	Residues 119-130							
Target-contacting positions:	118* (-..n-) 120* (-..n'-) 121* (-..n-) 124* (-..n-)							
Target sequence:	-nnn-							
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1aay chain A)					
			118*	119	120*	121*	123	124*
Subfamily 1								
ZF39_MOUSE	ZFZFP-39 (frag)	<i>Mus musculus</i>	a (R)	a (K)	c (S)	b (H)	h (G)	a (R)
TF3A_YEAST	TFIIIa	<i>S. cerevisiae</i>	a (K)	a (K)	c (S)	b (H)	e (E)	a (R)
Subfamily 2								
1meyF3	artificial	-	a (R)	c (S)	e (D)	b (H)	c (S)	a (R)
ZMS1_YEAST	ZFZms1	<i>S. cerevisiae</i>	a (R)	d (Q)	e (E)	b (H)	a (K)	a (R)
Subfamily 3								
MLZ4_MOUSE	ZFMlz-4	<i>Mus musculus</i>	a (R)	c (S)	c (S)	b (H)	h (A)	d (Q)
ZN19_HUMAN	ZF19	<i>Homo sapiens</i>	a (R)	g (T)	c (S)	b (H)	c (S)	d (Q)
Subfamily 4								
GLI4_HUMAN	Gli4	<i>Homo sapiens</i>	b (H)	c (S)	c (S)	b (H)	g (T)	d (Q)
ZN83_HUMAN	ZF83	<i>Homo sapiens</i>	b (H)	h (I)	c (S)	b (H)	h (A)	d (Q)
Subfamily 5								
ZFH1_DROME	Zinc-finger 1	<i>D. melanogaster</i>	b (H)	a (K)	b (H)	b (H)	g (T)	e (E)
NIL2_HUMAN	Nil-2-a ZF	<i>Homo sapiens</i>	b (H)	a (K)	b (H)	b (H)	h (I)	e (E)
Subfamily 6								
Z124_HUMAN	ZF124	<i>Homo sapiens</i>	a (R)	c (S)	c (S)	b (H)	a (R)	e (D)
ZG9_XENLA	Gastrula ZF Xlcf9.1	<i>Xenopus laevis</i>	a (R)	a (R)	c (S)	b (H)	i (M)	e (D)
Subfamily 7								
WT1_ALLMI	Wilms' tumor (frag)	<i>A. mississippiensis</i>	a (R)	c (S)	e (D)	b (H)	a (K)	g (T)
EGR2_CRILO	Egr-2	<i>C. longicaudatus</i>	a (R)	c (S)	e (D)	b (H)	g (T)	g (T)
1aayA2	Zif268	<i>Mus musculus</i>	a (R)	c (S)	e (D)	b (H)	g (T)	g (T)
Subfamily 8								
ZG3_XENLA	Gastrula ZF Xlcf3.1	<i>Xenopus laevis</i>	a (R)	d (N)	e (D)	b (H)	d (Q)	h (I)
ZXDA_HUMAN	ZF x-linked zxda (frag)	<i>Homo sapiens</i>	a (R)	h (A)	e (E)	b (H)	a (K)	h (G)
Subfamily 9								
Z126_HUMAN	ZF 126 (frag)	<i>Homo sapiens</i>	a (K)	b (H)	c (S)	b (H)	d (Q)	i (C)
ZN80_CERAE	ZF 80	<i>C. aethiops</i>	a (R)	a (R)	c (S)	b (H)	h (L)	i (C)
Subfamily 10								
Z135_HUMAN	ZF135	<i>Homo sapiens</i>	b (H)	c (S)	c (S)	c (S)	g (T)	a (K)
Z154_HUMAN	ZF154 (frag)	<i>Homo sapiens</i>	b (H)	d (N)	c (S)	c (S)	h (I)	a (K)
Subfamily 11								
2drpA1	Tramtrack	<i>D. melanogaster</i>	b (H)	h (I)	c (S)	d (N)	i (C)	a (R)
ZF38_MOUSE	ZFp-38 (ctfin51)	<i>Mus musculus</i>	b (H)	c (S)	c (S)	d (N)	d (N)	a (K)
Subfamily 12								
ZF37_MOUSE	ZFp-37	<i>Mus musculus</i>	b (H)	c (S)	c (S)	d (N)	i (M)	d (Q)
ZG53_XENLA	Gastrula ZF Xlcf53.1	<i>Xenopus laevis</i>	b (H)	b (H)	c (S)	d (N)	g (T)	d (N)
Subfamily 13								
Z134_HUMAN	ZF134	<i>Homo sapiens</i>	a (R)	a (K)	e (D)	d (N)	g (T)	d (Q)
MSN2_YEAST	ZFMsn2	<i>S. cerevisiae</i>	a (R)	c (S)	e (D)	d (N)	c (S)	d (Q)
MSN4_YEAST	ZFMsn4	<i>S. cerevisiae</i>	a (R)	c (S)	e (D)	d (N)	c (S)	d (Q)
RGM1_YEAST	repressor Rgm1	<i>S. cerevisiae</i>	a (R)	h (I)	e (D)	d (N)	a (R)	d (Q)
Subfamily 14								
TRA1_CAEEL	Sex-det. transformer 1	<i>C. elegans</i>	a (R)	h (L)	e (E)	d (N)	a (K)	g (T)
GLI1_HUMAN	ZF Gli1	<i>Homo sapiens</i>	a (R)	h (L)	e (E)	d (N)	a (K)	g (T)
Subfamily 15								
ZN41_HUMAN	ZF41 (frag)	<i>Homo sapiens</i>	b (H)	a (R)	g (T)	d (N)	g (T)	g (T)
ZN75_HUMAN	ZF75	<i>Homo sapiens</i>	b (H)	d (N)	g (T)	d (N)	b (H)	g (T)
Subfamily 16								
ZN36_HUMAN	ZF36	<i>Homo sapiens</i>	b (H)	c (S)	c (S)	d (N)	h (I)	h (L)
ZF36_HUMAN	ZF36	<i>Homo sapiens</i>	b (H)	f (F)	c (S)	d (N)	a (K)	h (V)
Subfamily 17								
2drpA2	Tramtrack	<i>D. melanogaster</i>	a (R)	a (K)	e (D)	d (N)	g (T)	h (A)

ZIC1_MOUSE	ZF Zic1	<i>Mus musculus</i>	a (R)	c (S)	e (E)	d (N)	a (K)	h (I)
Subfamily 18								
EGR1_BRARE	Egr-1	<i>Brachydanio rerio</i>	a (R)	c (S)	e (D)	e (E)	g (T)	a (R)
EGR2_BRARE	Egr-2	<i>Brachydanio rerio</i>	a (R)	c (S)	e (D)	e (E)	g (T)	a (R)
EKLF_HUMAN	Erythroid krueppel	<i>Homo sapiens</i>	a (R)	c (S)	e (D)	e (E)	g (T)	a (R)
ZN18_HUMAN	ZF18	<i>Homo sapiens</i>	a (R)	c (S)	c (S)	e (D)	h (V)	a (K)
SP1_HUMAN	TF Sp1	<i>Homo sapiens</i>	a (R)	c (S)	e (D)	e (E)	d (Q)	a (R)
SP2_HUMAN	TF Sp2	<i>Homo sapiens</i>	a (R)	c (S)	e (D)	e (E)	d (Q)	a (R)
1aayA1	Zif268	<i>Mus musculus</i>	a (R)	c (S)	e (D)	e (E)	g (T)	a (R)
1aayA3	Zif268	<i>Mus musculus</i>	a (R)	c (S)	e (D)	e (E)	a (K)	a (R)
EKLF_MOUSE	Erythroid krueppel	<i>Mus musculus</i>	a (R)	c (S)	e (D)	e (E)	g (T)	a (R)
Subfamily 19								
ZFA_MOUSE	ZF autosomal	<i>Mus musculus</i>	b (H)	f (P)	c (S)	e (E)	a (K)	a (K)
ZF92_MOUSE	ZFp-92 (frag)	<i>Mus musculus</i>	b (H)	c (S)	c (S)	e (D)	g (T)	a (K)
Subfamily 20								
ZF90_MOUSE	ZFp-90	<i>Mus musculus</i>	a (R)	c (S)	c (S)	h (A)	g (T)	a (K)
Z125_HUMAN	ZF125 (frag)	<i>Homo sapiens</i>	a (K)	i (C)	c (S)	h (L)	d (Q)	a (R)
Subfamily 21								
ZN12_HUMAN	ZF12	<i>Homo sapiens</i>	a (R)	h (L)	c (S)	h (Y)	g (T)	h (V)
ZG42_XENLA	Gastrula ZF Xlcf42.1	<i>Xenopus laevis</i>	a (R)	a (R)	c (S)	h (G)	g (T)	h (A)
Subfamily 22								
ZG20_XENLA	Gastrula ZF Xfg20-1	<i>Xenopus laevis</i>	b (H)	a (K)	c (S)	h (V)	a (K)	h (L)
Z11B_HUMAN	ZF11b (frag)	<i>Homo sapiens</i>	b (H)	a (K)	c (S)	h (A)	g (T)	h (L)
Subfamily 23								
Z138_HUMAN	ZF138 (frag)	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	b (H)	g (T)	a (R)
Z132_HUMAN	ZF132	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	b (H)	h (L)	a (R)
ZO71_XENLA	Oocyte ZF Xlcof7.1	<i>Xenopus laevis</i>	d (N)	d (Q)	c (S)	b (H)	h (A)	a (R)
Z117_HUMAN	ZF117	<i>Homo sapiens</i>	d (Q)	g (T)	c (S)	b (H)	h (I)	a (R)
Subfamily 24								
HF12_HUMAN	ZF12 (frag)	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	b (H)	h (Y)	d (Q)
ZKR1_CHICK	ZFCkr1	<i>Gallus gallus</i>	d (Q)	c (S)	c (S)	b (H)	h (V)	d (Q)
ZN22_HUMAN	ZF22	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	b (H)	h (I)	d (Q)
ZN38_HUMAN	ZF38	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	b (H)	h (Y)	d (Q)
Subfamily 25								
ZN25_HUMAN	ZF25	<i>Homo sapiens</i>	d (Q)	a (K)	c (S)	b (H)	g (T)	h (V)
Z177_HUMAN	ZF177	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	b (H)	d (N)	h (V)
ZO10_XENLA	Oocyte ZF Xlcof10	<i>Xenopus laevis</i>	d (Q)	a (K)	c (S)	b (H)	g (T)	h (A)
Subfamily 26								
1meyC1	artificial	-	d (Q)	c (S)	c (S)	d (N)	d (Q)	a (K)
OZF_BOVIN	ZF Ozf	<i>Bos taurus</i>	d (Q)	a (K)	c (S)	d (N)	h (I)	a (R)
ZN85_HUMAN	ZF85	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	d (N)	g (T)	a (K)
Subfamily 27								
KRUP_CUPSA	Krueppel (frag)	<i>Cupiennius salei</i>	d (Q)	h (V)	h (A)	d (N)	a (R)	a (R)
HELI_MOUSE	ZF helios	<i>Mus musculus</i>	d (Q)	a (K)	h (G)	d (N)	h (L)	a (R)
KRUP_TRICA	Krueppel (frag)	<i>T. castaneum</i>	d (Q)	h (V)	h (A)	d (N)	a (R)	a (R)
Subfamily 28								
ZN84_HUMAN	ZF84	<i>Homo sapiens</i>	d (Q)	a (K)	c (S)	d (Q)	g (T)	c (S)
ZG8_XENLA	Gastrula ZF Xlcf8.2db	<i>Xenopus laevis</i>	d (Q)	a (K)	c (S)	d (N)	h (V)	c (S)
Subfamily 29								
ZN43_HUMAN	ZF 43	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	d (N)	g (T)	g (T)
ZF27_MOUSE	ZFp-27 (mkr4)	<i>Mus musculus</i>	d (N)	a (R)	c (S)	d (N)	h (I)	g (T)
ZNG1_RAT	ZF Gfi-1	<i>Rattus norvegicus</i>	d (Q)	c (S)	c (S)	d (N)	h (I)	g (T)
ZG49_XENLA	Gastrula ZF Xlcf49.1	<i>Xenopus laevis</i>	d (Q)	a (K)	c (S)	d (N)	d (Q)	g (T)
Subfamily 30								
Z33B_HUMAN	ZF 33b (frag)	<i>Homo sapiens</i>	d (Q)	a (K)	c (S)	d (N)	h (I)	h (V)
ZF13_MOUSE	ZFp-13 (krox-8)	<i>Mus musculus</i>	d (Q)	a (R)	c (S)	d (N)	h (I)	h (A)
Subfamily 31								
Z11A_HUMAN	ZF11a	<i>Homo sapiens</i>	d (Q)	a (K)	c (S)	e (D)	g (T)	a (K)
1meyC2	artificial	-	d (Q)	c (S)	c (S)	e (D)	d (Q)	a (K)
Subfamily 32								
MOK2_HUMAN	ZFmok-2	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	e (D)	a (R)	h (I)
ZF35_MOUSE	ZFp-35	<i>Mus musculus</i>	d (Q)	c (S)	c (S)	e (D)	i (M)	h (I)
Subfamily 33								
ZN91_HUMAN	ZF91	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	g (T)	g (T)	a (R)
ZN15_HUMAN	ZF15	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	g (T)	i (M)	a (K)
ZO72_XENLA	Oocyte ZF Xlcof7.2	<i>Xenopus laevis</i>	d (N)	d (Q)	c (S)	g (T)	h (A)	a (R)
Subfamily 34								

Z133_HUMAN	ZF133	<i>Homo sapiens</i>	d (Q)	a (K)	c (S)	h (A)	h (V)	a (R)
XFIN_XENLA	Xfin	<i>Xenopus laevis</i>	d (Q)	c (S)	c (S)	h (A)	h (V)	a (K)
Subfamily 35								
ZG71_XENLA	Gastrula ZF Xlcf71.1	<i>Xenopus laevis</i>	e (D)	a (R)	c (S)	b (H)	g (T)	a (R)
1ubdC2	YY1	<i>Homo sapiens</i>	e (E)	c (S)	c (S)	a (K)	a (K)	a (R)
Subfamily 36								
SLUG_XENLA	Slug (Xslu)	<i>Xenopus laevis</i>	e (D)	a (R)	c (S)	d (N)	a (R)	h (A)
SNAI_DROME	Snail	<i>D. melanogaster</i>	e (D)	a (R)	c (S)	d (N)	a (R)	h (A)
Subfamily 37								
1ubdC1	YY1	<i>Homo sapiens</i>	e (D)	d (N)	c (S)	h (A)	a (R)	a (K)
ZO6_XENLA	Oocyte ZF Xlcof6	<i>Xenopus laevis</i>	e (E)	a (K)	c (S)	h (I)	d (Q)	a (K)
Subfamily 38								
TF3A_BUFAM	TFIIIA	<i>Bufo americanus</i>	g (T)	h (L)	b (H)	b (H)	d (N)	a (R)
MFG2_MOUSE	ZF Mfg-2 (frag)	<i>Mus musculus</i>	g (T)	h (A)	a (R)	b (H)	h (V)	a (K)
Subfamily 39								
ZG66_XENLA	Gastrula ZF Xlcf66.1	<i>Xenopus laevis</i>	h (Y)	a (R)	h (A)	b (H)	h (V)	a (R)
ZN23_HUMAN	ZF23	<i>Homo sapiens</i>	h (I)	d (N)	h (A)	a (K)	g (T)	a (R)
Subfamily40								
EVI1_MOUSE	Ecotropic virus int-1	<i>Mus musculus</i>	h (I)	c (S)	c (S)	d (N)	d (Q)	a (R)
ZO14_XENLA	Oocyte ZF Xlcof14	<i>Xenopus laevis</i>	h (A)	a (K)	d (Q)	d (N)	a (K)	a (R)
Subfamily 41								
1ubdC3	YY1	<i>Homo sapiens</i>	h (L)	e (D)	f (F)	d (N)	a (R)	g (T)
TY1_HUMAN	Repressor YY1	<i>Homo sapiens</i>	h (L)	e (D)	f (F)	d (N)	a (K)	g (T)
REX1_MOUSE	Rex-1	<i>Mus musculus</i>	h (L)	e (D)	f (F)	d (N)	a (R)	g (T)
TY1_MOUSE	Repressor YY1	<i>Mus musculus</i>	h (L)	e (D)	f (F)	d (N)	a (R)	g (T)
Subfamily 42								
KRXC_MOUSE	Krox-6.1b- (frag)	<i>Mus musculus</i>	i (C)	c (S)	c (S)	h (Y)	c (S)	a (K)
KRXA_MOUSE	Krox-6.1a (frag)	<i>Mus musculus</i>	i (C)	c (S)	c (S)	h (Y)	g (T)	a (K)
Individual proteins								
Z140_HUMAN	ZF140	<i>Homo sapiens</i>	a (R)	h (A)	c (S)	d (N)	g (T)	a (R)
ZN13_HUMAN	ZF13	<i>Homo sapiens</i>	a (R)	h (A)	c (S)	d (N)	h (L)	h (A)
ZFY1_MOUSE	ZF Y-chromosomal 1	<i>Mus musculus</i>	b (H)	f (P)	c (S)	h (A)	a (K)	a (K)
ZN16_HUMAN	ZF16	<i>Homo sapiens</i>	b (H)	c (S)	c (S)	h (A)	h (I)	d (Q)
ZN30_HUMAN	ZF30	<i>Homo sapiens</i>	a (R)	h (A)	c (S)	h (Y)	h (V)	d (Q)
Z165_HUMAN	ZF165	<i>Homo sapiens</i>	b (H)	c (S)	c (S)	a (K)	h (A)	a (R)
KRUH_DROME	Kruempel ZF (frag)	<i>D. melanogaster</i>	b (H)	c (S)	h (G)	a (K)	b (H)	a (R)
TF3A_HUMAN	TFIIIA	<i>Homo sapiens</i>	a (R)	e (D)	h (Y)	b (H)	c (S)	a (R)
MAZ_HUMAN	Myc-associated ZF	<i>Homo sapiens</i>	a (R)	a (K)	e (D)	a (R)	c (S)	h (Y)
KRUP_APIME	Kruempel (frag)	<i>Apis mellifera</i>	a (R)	e (D)	b (H)	b (H)	a (K)	g (T)
ZG16_XENLA	Gastrula ZF Xlcf16.1	<i>Xenopus laevis</i>	a (R)	a (R)	c (S)	b (H)	d (N)	c (S)
ZG32_XENLA	Gastrula ZF Xlcf32.1	<i>Xenopus laevis</i>	a (R)	a (K)	c (S)	a (K)	a (K)	g (T)
ZN10_HUMAN	ZF10	<i>Homo sapiens</i>	a (R)	c (S)	c (S)	b (H)	h (I)	h (G)
ZN27_HUMAN	ZF27	<i>Homo sapiens</i>	b (H)	e (D)	c (S)	d (Q)	d (Q)	e (E)
ZO84_XENLA	Oocyte ZF Xlcof8.4	<i>Xenopus laevis</i>	a (R)	c (S)	c (S)	e (D)	d (N)	h (V)
ZN31_HUMAN	ZF31	<i>Homo sapiens</i>	a (K)	c (S)	c (S)	g (T)	h (A)	d (N)
HMS1_YEAST	ZF Hms1	<i>S. cerevisiae</i>	a (R)	a (K)	c (S)	i (W)	a (K)	a (R)
BCL6_HUMAN	B-cell lymphoma 6	<i>Homo sapiens</i>	b (H)	h (L)	d (Q)	g (T)	a (K)	c (S)
ZO20_XENLA	Oocyte ZF Xlcof20	<i>Xenopus laevis</i>	a (K)	d (N)	e (D)	h (V)	h (L)	h (I)
ZF29_MOUSE	ZFp-29	<i>Mus musculus</i>	a (R)	c (S)	f (P)	d (N)	h (I)	h (A)
ZN08_HUMAN	ZF8	<i>Homo sapiens</i>	b (H)	c (S)	g (T)	b (H)	g (T)	h (V)
Z141_HUMAN	ZF141	<i>Homo sapiens</i>	a (R)	c (S)	g (T)	g (T)	g (T)	a (K)
ZN20_HUMAN	ZF20	<i>Homo sapiens</i>	a (R)	c (S)	g (T)	g (T)	f (P)	h (V)
ZF64_HUMAN	ZF647 (frag)	<i>Homo sapiens</i>	b (H)	c (S)	g (T)	h (V)	a (R)	c (S)
MFG3_MOUSE	ZF Mfg-3	<i>Mus musculus</i>	a (R)	c (S)	g (T)	h (G)	a (R)	h (I)
Z131_HUMAN	ZF131 (frag)	<i>Homo sapiens</i>	b (H)	f (F)	h (G)	b (H)	a (K)	e (E)
ZFP1_MOUSE	ZFp-1 (mkr1)	<i>Mus musculus</i>	b (H)	a (K)	h (A)	d (N)	h (I)	a (K)
AZF1_YEAST	Asparagine-rich ZF	<i>S. cerevisiae</i>	a (R)	a (K)	h (G)	d (N)	h (A)	h (A)
IA1_HUMAN	ZF Ia-1	<i>Homo sapiens</i>	a (R)	d (Q)	h (A)	h (Y)	a (R)	a (K)
FZF1_YEAST	ZF Fzf1	<i>S. cerevisiae</i>	a (R)	f (P)	i (C)	b (H)	a (R)	h (V)
TF3A_XENBO	TFIIIA	<i>Xenopus borealis</i>	c (S)	h (L)	b (H)	b (H)	g (T)	a (R)
ZN46_HUMAN	ZF46	<i>Homo sapiens</i>	c (S)	a (R)	c (S)	d (N)	a (R)	d (Q)
Z136_HUMAN	ZF136	<i>Homo sapiens</i>	c (S)	c (S)	g (T)	c (S)	a (R)	h (I)
ZG64_XENLA	Gastrula ZF Xlcf64.1	<i>Xenopus laevis</i>	d (Q)	a (K)	a (K)	h (A)	a (R)	a (R)
Z195_HUMAN	ZF195	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	b (H)	c (S)	e (E)
ZN29_HUMAN	ZF29	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	c (S)	g (T)	d (Q)
ZN26_HUMAN	ZF26	<i>Homo sapiens</i>	d (Q)	a (K)	c (S)	c (S)	c (S)	e (E)
ZN35_HUMAN	ZF35	<i>Homo sapiens</i>	d (Q)	a (R)	c (S)	c (S)	g (T)	h (V)
Z184_HUMAN	ZF184 (frag)	<i>Homo sapiens</i>	d (Q)	b (H)	c (S)	d (N)	g (T)	d (Q)
AEF1_DROME	Adult enhancer factor 1	<i>D. melanogaster</i>	d (Q)	c (S)	c (S)	g (T)	g (T)	d (N)
ZN90_HUMAN	ZF90	<i>Homo sapiens</i>	d (Q)	c (S)	c (S)	g (T)	h (A)	g (T)
ZO28_XENLA	Oocyte ZF Xlcof28	<i>Xenopus laevis</i>	d (Q)	d (Q)	c (S)	g (T)	h (V)	h (V)
Z191_HUMAN	ZF191	<i>Homo sapiens</i>	d (Q)	d (N)	c (S)	h (G)	h (I)	d (N)
ZN45_HUMAN	ZF45 (brc1744)	<i>Homo sapiens</i>	d (Q)	a (R)	c (S)	h (Y)	d (Q)	h (A)
ZO22_XENLA	Oocyte ZF Xlcof22	<i>Xenopus laevis</i>	d (N)	d (Q)	c (S)	i (C)	a (R)	h (V)
ZG48_XENLA	Gastrula ZF Xlcf48.2	<i>Xenopus laevis</i>	d (Q)	c (S)	f (P)	d (Q)	e (D)	h (L)

REQN_RAT	Zinc-finger neuro-D4	<i>Rattus norvegicus</i>	d (N)	a (R)	f (P)	h (G)	c (S)	h (Y)
ZN74_HUMAN	ZF74	<i>Homo sapiens</i>	d (Q)	a (R)	g (T)	b (H)	g (T)	a (R)
ZF28_MOUSE	ZFp-28 (mkr5)	<i>Mus musculus</i>	d (Q)	g (T)	g (T)	b (H)	h (I)	d (Q)
1ubdC4	YY1	<i>Homo sapiens</i>	d (Q)	c (S)	g (T)	d (N)	a (K)	c (S)
ZG46_XENLA	Gastrula ZF Xlcfg46.1	<i>Xenopus laevis</i>	d (Q)	a (K)	g (T)	d (N)	d (N)	g (T)
ZG57_XENLA	Gastrula ZF Xlcfg57.1	<i>Xenopus laevis</i>	d (Q)	a (K)	g (T)	d (N)	h (L)	i (C)
ZN42_HUMAN	ZF42	<i>Homo sapiens</i>	d (Q)	a (R)	h (L)	a (K)	g (T)	a (R)
ZK23_HUMAN	ZF kox23 (frag)	<i>Homo sapiens</i>	d (Q)	c (S)	h (A)	c (S)	h (I)	d (Q)
HKR3_HUMAN	Krueppel-related ZF3	<i>Homo sapiens</i>	d (Q)	a (K)	h (A)	d (N)	d (N)	i (M)
Z123_HUMAN	ZF123 (frag)	<i>Homo sapiens</i>	d (Q)	a (K)	h (I)	g (T)	h (I)	d (Q)
ZG5A_XENLA	Gastrula ZF Xlcfg51.1a	<i>Xenopus laevis</i>	d (Q)	a (R)	i (M)	b (H)	h (I)	e (E)
ZO61_XENLA	Oocyte ZF Xlcof6.1	<i>Xenopus laevis</i>	d (Q)	c (S)	i (M)	d (Q)	h (I)	a (R)
KR2_MOUSE	Mkr2 (ZF2)	<i>Mus musculus</i>	d (Q)	c (S)	i (M)	d (N)	g (T)	h (V)
ZG62_XENLA	Gastrula ZF Xlcfg62.1	<i>Xenopus laevis</i>	e (E)	a (K)	a (R)	g (T)	a (K)	b (H)
ZG26_XENLA	Gastrula ZF Xlcfg26.1	<i>Xenopus laevis</i>	e (E)	a (K)	a (K)	g (T)	a (R)	e (E)
HKR2_HUMAN	Krueppel-related ZF2	<i>Homo sapiens</i>	e (E)	c (S)	c (S)	c (S)	h (A)	a (K)
ZN81_HUMAN	ZF81 (frag)	<i>Homo sapiens</i>	e (D)	a (R)	c (S)	d (N)	d (N)	a (K)
ZG7_XENLA	Gastrula ZF Xlcfg7.1	<i>Xenopus laevis</i>	e (D)	a (K)	c (S)	d (N)	a (R)	c (S)
Z37A_HUMAN	ZF37a	<i>Homo sapiens</i>	e (E)	a (K)	c (S)	g (T)	g (T)	a (K)
ZN21_HUMAN	ZF21	<i>Homo sapiens</i>	e (E)	a (K)	c (S)	g (T)	g (T)	h (V)
ZN80_HUMAN	ZF80	<i>Homo sapiens</i>	e (E)	a (K)	h (V)	e (D)	h (V)	a (R)
Z157_HUMAN	ZF157	<i>Homo sapiens</i>	e (E)	a (K)	h (A)	g (T)	g (T)	h (I)
Z151_HUMAN	ZF151 (miz-1)	<i>Homo sapiens</i>	e (D)	f (P)	h (G)	h (A)	d (Q)	a (R)
ZG44_XENLA	Gastrula ZF Xlcfg44.2	<i>Xenopus laevis</i>	e (D)	a (R)	h (I)	h (I)	d (Q)	h (A)
ZN17_HUMAN	ZF17	<i>Homo sapiens</i>	e (D)	c (S)	i (C)	g (T)	a (K)	c (S)
ZN14_HUMAN	ZF14	<i>Homo sapiens</i>	f (F)	c (S)	c (S)	c (S)	d (Q)	a (R)
ZO2_XENLA	Oocyte ZF Xlcof2	<i>Xenopus laevis</i>	f (F)	g (T)	g (T)	c (S)	h (I)	a (R)
ZN44_HUMAN	ZF44	<i>Homo sapiens</i>	g (T)	c (S)	c (S)	c (S)	a (R)	a (K)
MFG1_MOUSE	ZF Mfg-1	<i>Mus musculus</i>	g (T)	c (S)	c (S)	d (N)	c (S)	e (E)
Z155_HUMAN	ZF155 (frag)	<i>Homo sapiens</i>	g (T)	a (K)	f (F)	d (N)	e (D)	h (L)
TF3A_RANPI	TFIIIA	<i>Rana pipiens</i>	g (T)	h (L)	f (F)	b (H)	g (T)	a (R)
CTCF_HUMAN	Repressor Ctfc	<i>Homo sapiens</i>	g (T)	h (V)	g (T)	h (L)	a (R)	d (N)
ZF26_MOUSE	ZFp-26 (mkr3)	<i>Mus musculus</i>	g (T)	c (S)	c (S)	h (G)	h (V)	e (E)
Z143_HUMAN	ZF143	<i>Homo sapiens</i>	g (T)	c (S)	d (N)	h (I)	a (K)	h (V)
MTF1_MOUSE	TF Mtf-1	<i>Mus musculus</i>	h (A)	c (S)	b (H)	b (H)	a (K)	g (T)
PLZF_HUMAN	ZF plzf	<i>Homo sapiens</i>	h (L)	a (K)	b (H)	d (Q)	e (E)	g (T)
ZN07_HUMAN	ZF7	<i>Homo sapiens</i>	h (L)	c (S)	c (S)	a (K)	h (I)	d (Q)
ZG29_XENLA	Gastrula ZF Xlcfg29.1	<i>Xenopus laevis</i>	h (L)	a (K)	c (S)	a (R)	h (I)	h (A)
ZN80_GORGO	ZF80	<i>Gorilla gorilla</i>	h (Y)	d (N)	c (S)	c (S)	g (T)	a (R)
ZG52_XENLA	Gastrula ZF Xlcfg52.1	<i>Xenopus laevis</i>	h (V)	a (K)	c (S)	c (S)	h (L)	c (S)
ZF14_MOUSE	ZFp-14 (krox-9)	<i>Mus musculus</i>	h (L)	h (L)	c (S)	d (Q)	g (T)	d (Q)
OZF_MOUSE	ZF Ozf	<i>Mus musculus</i>	h (G)	a (K)	c (S)	d (N)	g (T)	e (E)
ZO29_XENLA	Oocyte ZF Xlcof29	<i>Xenopus laevis</i>	h (Y)	a (R)	c (S)	d (N)	i (M)	h (V)
ZG17_XENLA	Gastrula ZF Xlcfg17.1	<i>Xenopus laevis</i>	h (A)	c (S)	c (S)	e (D)	a (R)	h (V)
ZG67_XENLA	Gastrula ZF Xlcfg67.1	<i>Xenopus laevis</i>	h (Y)	a (R)	c (S)	h (V)	i (M)	e (E)
ZO26_XENLA	Oocyte ZF Xlcof26	<i>Xenopus laevis</i>	h (V)	a (K)	d (N)	c (S)	a (R)	a (K)
ZG58_XENLA	Gastrula ZF Xlcfg58.1	<i>Xenopus laevis</i>	h (L)	a (K)	d (Q)	h (L)	g (T)	c (S)
ZG5_XENLA	Gastrula ZF5-1	<i>Xenopus laevis</i>	h (L)	a (K)	e (D)	c (S)	b (H)	a (R)
KID1_RAT	Renal TF Kid-1	<i>Rattus norvegicus</i>	h (L)	c (S)	g (T)	c (S)	h (Y)	a (K)
ZN32_HUMAN	ZF32	<i>Homo sapiens</i>	h (A)	a (K)	h (A)	d (N)	h (V)	g (T)
ZN12_MICSA	ZFmsa12a	<i>M. salmoides</i>	h (V)	c (S)	h (G)	d (N)	d (N)	h (I)
ZN39_HUMAN	ZF39	<i>Homo sapiens</i>	i (W)	c (S)	c (S)	a (K)	g (T)	e (E)
KID1_MOUSE	Renal TF Kid-1	<i>Mus musculus</i>	i (C)	d (N)	c (S)	c (S)	c (S)	d (N)
HKR1_HUMAN	Krueppel-related ZF1	<i>Homo sapiens</i>	i (W)	a (K)	c (S)	d (N)	a (K)	g (T)
ZG28_XENLA	Gastrula ZF Xlcfg28.1	<i>Xenopus laevis</i>	i (C)	d (N)	c (S)	d (Q)	d (N)	h (L)
Z174_HUMAN	ZF174	<i>Homo sapiens</i>	i (W)	d (N)	c (S)	e (E)	a (K)	a (R)
TF3A_ICTPU	TFIIIA	<i>I. punctatus</i>	i (M)	e (E)	h (G)	c (S)	a (K)	a (R)
Z142_HUMAN	ZF142	<i>Homo sapiens</i>	i (W)	h (A)	h (A)	h (G)	a (R)	b (H)
Conservation score			56.5	60.2	65.5	58.2	53.5	54.1

Table A1.20. Hormone receptor family (Multi-specific)

The hormone receptors modulate transcription of numerous regulatory pathways as a response to hormone binding. Proteins use an α -helix from a zinc-coordinating motif to bind in the DNA major groove.

Five positions from the motif interact with the target sequence. There are three conserved positions that are common to all family members. Position 324 comprises lysine that binds guanine in a bidentate interaction (-.G....-). Position 329 is an arginine that also binds guanine in a bidentate interaction (-....G'.-). Position 328 consists of arginine or lysine which binds either guanine and thymine (-.GT..-) or two adenines (-..AA..-) in a complex interaction. Experimental analysis shows that amino acid substitutions at these positions abolish *in vivo* binding by the glucocorticoid receptor⁵³.

Interactions from positions 321 and 325 allow discrimination between the two hormone response elements by the members of the two subfamilies. The half-site sequences only differ at two base-steps. At position 321, glutamate in subfamily 1 accepts a hydrogen bond from cytosine (-.C'...-) while glycine in subfamily 2 does not interact with any base. At position 325 glycine or alanine in subfamily 1 does not interact but valine in subfamily 2 recognises thymine through methyl-methyl contact (-...T'..-). Mutations at positions 321, 322 (non-interacting), 325, in the glucocorticoid receptor (11at) are sufficient to swap specificity for the half-site sequence to that of the oestrogen receptor⁵⁴. Amino acids of orthologous proteins from different organisms are conserved and the proteins belong to the same subfamilies.

Hormone receptors function as homo- and heterodimers, which allows recognition of different combinations of hormone response elements. Specificity for the full target sequence also depends on the spacing and relative orientations of the half-site, which is recognised by the type of dimers that are formed⁵⁵.

Point mutations, especially in the DNA-binding, dimerisation and co-factor-binding domains of hormone receptors cause a wide range of disorders such as growth, rickets and thyroid gland disorders⁵⁶.

Table A1.20

Hormone receptor family (Multi-specific)										
Representative structure:	2nll Chain B									
Core DNA-binding region:	Residues 310-335									
Target-contacting positions:	321 (-..n'...-) 324* (-G....-) 325 (-..n'...-) 328* (-..nn...-) 329* (-...G'-)									
Target sequence:	-AGGTCA- (subfamily 1) -AGAACA- (subfamily 2)									
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 2nll chain B)							
			321	324*	325	328*	329*	375	376	384
Subfamily 1										
ECR_AEDAE	Ecdysone R	Aedes aegypti	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	h (V)	a (K)
CF1_BOMMO	Rxr type hormone R Cf1	Bombyx mori	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	d (Q)	- (-)
ECR_BOMMO	Ecdysone R	Bombyx mori	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	- (-)	- (-)
FTF1_BOMMO	Hormone R Ftz-F1	Bombyx mori	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	a (R)	a (R)
VDR_BOVIN	Vitamin d3 R	Bos taurus	e (E)	a (K)	h (G)	a (R)	a (R)	h (I)	h (L)	a (K)
AD4B_BOVIN	Hormone R Ad4bp	Bos taurus	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	a (R)	a (R)
CNR8_CAEEL	Hormone R Cnr8	Caenorhabditis elegans	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	a (R)	- (-)
CNRD_CAEEL	Hormone R Cnr14	Caenorhabditis elegans	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	a (R)	i (M)
NHR2_CAEEL	Hormone R Nhr-2	Caenorhabditis elegans	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	a (R)	- (-)
NGF1_CANFA	Orphan nuclear R Ngfi-b	Canis familiaris	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	a (R)	- (-)
PPAR_CAVPO	Peroxisome proliferator activated al	Cavia porcellus	e (E)	a (K)	h (G)	a (R)	a (R)	h (I)	- (-)	- (-)
ECR_DROME	Ecdysone R	Drosophila melanogaster	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	h (V)	a (K)
E75A_DROME	Ecdysone-inducible protein E75-a	Drosophila melanogaster	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	- (-)	a (K)
FTFB_DROME	Hormone R Ftz-F1 β (dhr39)	Drosophila melanogaster	e (E)	a (K)	h (G)	a (K)	a (R)	h (I)	- (-)	- (-)
HR78_DROME	Hormone R hr78 (dhr78)	Drosophila melanogaster	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	d (Q)	e (D)
HR96_DROME	Hormone R hr96 (dhr96)	Drosophila melanogaster	e (E)	a (K)	h (A)	a (R)	a (R)	h (I)	i (M)	a (K)
ESTR_CHICK	Oestrogen R	Gallus gallus	e (E)	a (K)	h (A)	a (K)	a (R)	h (I)	a (R)	- (-)
RRXA_CHICK	Retinoic acid R Rxr- α	Gallus gallus	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	- (-)	- (-)
1hcqA	Oestrogen R	Homo sapiens	e (E)	a (K)	h (A)	a (K)	a (R)	- (-)	- (-)	- (-)
1hcqB	Oestrogen R	Homo sapiens	e (E)	a (K)	h (A)	a (K)	a (R)	- (-)	- (-)	- (-)
2nllA	Retinoic acid R	Homo sapiens	e (E)	a (K)	h (G)	a (K)	a (R)	- (-)	- (-)	- (-)
2nllB	Thyroid hormone R	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	h (L)	a (K)
ERR1_HUMAN	Hormone R err1 (estrogen-like)	Homo sapiens	e (E)	a (K)	h (A)	a (K)	a (R)	h (V)	a (R)	- (-)
ERR2_HUMAN	Hormone R err2 (estrogen-like)	Homo sapiens	e (E)	a (K)	h (A)	a (K)	a (R)	h (V)	a (R)	- (-)
HNF4_HUMAN	Hepatocyte nuclear factor 4	Homo sapiens	e (D)	a (K)	h (G)	a (R)	a (R)	h (V)	d (Q)	- (-)
NER_HUMAN	Nuclear R Ner	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	h (L)	- (-)
NOT_HUMAN	Immediate-early response protein not	Homo sapiens	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	a (R)	- (-)
PPAS_HUMAN	Peroxisome proliferator activated be	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (I)	- (-)	- (-)
PPAT_HUMAN	Peroxisome proliferator activated R	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (I)	- (-)	- (-)
ROR1_HUMAN	Nuclear R Ror- α -1	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	a (K)	- (-)
ROR4_HUMAN	Nuclear R Ror- α -4	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	a (K)	- (-)
RR1_HUMAN	Retinoic acid R α -1	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	a (R)	- (-)
RRXB_HUMAN	Retinoic acid R Rxr- β	Homo sapiens	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	d (Q)	- (-)
RRXG_HUMAN	Retinoic acid R Rxr- γ	Homo sapiens	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	d (Q)	- (-)
THA1_HUMAN	Thyroid hormone R α -1 (Ear-7-1)	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	h (L)	a (K)
THB1_HUMAN	Thyroid hormone R β -1	Homo sapiens	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	h (L)	a (K)
TR2_HUMAN	Orphan R	Homo sapiens	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	- (-)	e (E)
TR4_HUMAN	Orphan R r4 (Tak1)	Homo sapiens	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	- (-)	e (E)
E75B_MANSE	Ecdysone-inducible protein E75-b	Manduca sexta	e (D)	a (K)	h (G)	a (R)	a (R)	h (V)	- (-)	a (K)
ECR_MANSE	Ecdysone R	Manduca sexta	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	h (V)	- (-)
ROR4_MOUSE	Nuclear R Ror- α -4	Mus musculus	e (E)	a (K)	h (G)	a (R)	a (R)	h (V)	a (K)	d (Q)
ESTR_OREAU	Oestrogen R	Oreochromis aureus	e (E)	a (K)	h (A)	a (K)	a (R)	i (M)	a (R)	a (R)
ESTR_ORYLA	Oestrogen R	Oryzias latipes	e (E)	a (K)	h (A)	a (K)	a (R)	h (V)	a (R)	- (-)
1latA	Glucocorticoid R (mutant)	Rattus norvegicus	e (E)	a (K)	h (A)	a (K)	a (R)	- (-)	- (-)	- (-)
1latB	Glucocorticoid R (mutant)	Rattus norvegicus	e (E)	a (K)	h (A)	a (K)	a (R)	a (R)	- (-)	- (-)
ESTR_SALIR	Oestrogen R	Salmo irideus	e (E)	a (K)	h (A)	a (K)	a (R)	h (V)	a (K)	- (-)
NGF1_XENLA	Nerve growth factor protein I- β	Xenopus laevis	e (E)	a (K)	h (G)	a (K)	a (R)	h (V)	a (R)	- (-)
PPAR_XENLA	Peroxisome proliferator activated α	Xenopus laevis	e (E)	a (K)	h (G)	a (R)	a (R)	h (I)	- (-)	- (-)
PPAS_XENLA	Peroxisome proliferator activated β	Xenopus laevis	e (E)	a (K)	h (G)	a (R)	a (R)	h (I)	- (-)	- (-)
PPAT_XENLA	Peroxisome proliferator activated γ	Xenopus laevis	e (E)	a (K)	h (G)	a (R)	a (R)	h (I)	- (-)	- (-)
Subfamily 2										
GCR_AOTNA	Glucocorticoid R	Aotus nancymaee	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	a (K)	a (K)
PRGR_CHICK	Progesterone R	Gallus gallus	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	a (K)	- (-)
ANDR_HUMAN	Androgen R	Homo sapiens	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	a (K)	d (N)
MCR_HUMAN	Mineralocorticoid R	Homo sapiens	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	- (-)	a (K)
GCR_ONCMY	Glucocorticoid R	Oncorhynchus mykiss	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	a (K)	- (-)
GCR_SHEEP	Glucocorticoid R (fragment)	Ovis aries	h (G)	a (K)	h (V)	a (K)	a (R)	- (-)	- (-)	- (-)
1gluB	Glucocorticoid R	Rattus norvegicus	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	a (K)	- (-)
MCR_XENLA	Mineralocorticoid R (fragment)	Xenopus laevis	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	- (-)	a (K)
Individual proteins										
CSR1_CAEEL	Hormone R Csr-1	Caenorhabditis elegans	d (N)	a (K)	g (T)	a (R)	a (R)	h (I)	d (Q)	h (Y)
GCR_CAVPO	Glucocorticoid R	Cavia porcellus	h (G)	a (K)	h (V)	a (K)	a (R)	a (R)	a (K)	- (-)
7UP1_DROME	Steroid R seven-up type 1	Drosophila melanogaster	e (E)	a (K)	c (S)	a (K)	a (R)	- (-)	- (-)	- (-)
E75B_DROME	Ecdysone-inducible protein E75-b	Drosophila melanogaster	- (-)	- (-)	- (-)	a (R)	a (R)	h (V)	- (-)	a (K)
EGON_DROME	Embryonic gonad protein	Drosophila melanogaster	e (E)	a (K)	c (S)	h (G)	a (R)	- (-)	- (-)	- (-)
Conservation score			78.1	96.9	69.7	81.3	100	55.4	21.8	8.2

Table A1.21. GAL4 family (Multi-specific)

The family contains proteins that use the Gal4 zinc-coordinating motif and are involved in regulation of a wide range of metabolic pathways.

Two positions in the complex structure 1d66 contact the target sequence. Basic residues at position 17 (subfamily 1) use the peptide backbone to accept a bond from cytosine (-C.....-). Another basic residue at position 18 uses the peptide backbone to accept bonds from two cytosines (-CC'.....-) and the side chain as a donor to two guanines (-.GG.....-).

Only position 17 undergoes mutations in the family: hydrophobic residues are found in subfamily 2 and serine in subfamily 3. However, the importance of the amino acid type is unclear as the interaction is made through the peptide backbone. Experimental analysis of the target site for GAL4 shows that mutations at the first three base-steps of the half-site leads to loss of function ⁵⁷, but substitutions of the remaining base-steps does not affect binding.

Despite the apparent wide range of functions held by the proteins represented in the family, the amino acids at the interacting positions are remarkably well conserved. This is because target sequences are differentiated by recognising the spacing between the -CGG- triplets at either end of the site, rather than direct read-out of the bases. Specificity is therefore dependent on the length of the loop region which joins the DNA-binding and dimerisation domains and the method of dimerisation ⁵⁸. In a variation, the recently solved structure of the homologous HAP1 protein shows the dimer (not included in dataset) bound to directly repeated -CGG- triplets, therefore increasing the number of target sites that can be bound ⁵⁹.

Table A1.21

GAL4 family (Multi-specific)					
Representative structure:	1d66 chain A				
Core DNA-binding region:	Residues 8-35				
Target-contacting positions:	17 (-C.....-) 18* (-CGG.....-)				
Target sequence:	-CGGnnnnn- (variable separation between half-sites)				
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1d66 chain A)		
			17	18*	19
Subfamily 1					
UAY_EMENI	Positive purine utilisation regulator	<i>Emericella nidulans</i>	a (R)	a (K)	d (N)
AC15_NEUCR	Transcriptional activator Acu-15	<i>Neurospora crassa</i>	a (K)	a (K)	h (I)
FLUF_NEUCR	Conidial development protein fluffy	<i>Neurospora crassa</i>	a (K)	a (R)	g (T)
125d	Cd2-Gal4 protein	<i>Saccharomyces cerevisiae</i>	a (K)	a (K)	h (L)
1d66A	Gal4 protein	<i>Saccharomyces cerevisiae</i>	a (K)	a (K)	h (L)
1pyc	Cyp1 protein	<i>Saccharomyces cerevisiae</i>	a (R)	a (K)	h (V)
GAL4_YEAST	Gal4 protein	<i>Saccharomyces cerevisiae</i>	a (K)	a (K)	h (L)
MA1R_YEAST	Maltose fermentation regulator Mal1r	<i>Saccharomyces cerevisiae</i>	a (R)	a (R)	h (V)
MA3R_YEAST	Maltose fermentation regulator Mal3r	<i>Saccharomyces cerevisiae</i>	a (R)	a (R)	h (V)
MA6R_YEAST	Maltose fermentation regulator Mal6r	<i>Saccharomyces cerevisiae</i>	a (R)	a (R)	h (V)
PDR1_YEAST	Pleiotropic drug resistance	<i>Saccharomyces cerevisiae</i>	a (R)	a (K)	h (I)
PPR1_YEAST	Pyrimidine pathway regulator 1	<i>Saccharomyces cerevisiae</i>	a (K)	a (K)	h (I)
PUT3_YEAST	Proline utilization transactivator	<i>Saccharomyces cerevisiae</i>	a (R)	b (H)	h (I)
SIP4_YEAST	Sip4 protein	<i>Saccharomyces cerevisiae</i>	a (K)	a (K)	h (I)
THI1_SCHPO	Thiamine repressible genes regulator	<i>Schizosaccharomyces pombe</i>	a (K)	a (K)	h (I)
LAC9_KLULA	Lactose regulatory protein Lac9	<i>Kluyveromyces lactis</i>	a (K)	a (K)	i (W)
Subfamily 2					
QA1F_NEUCR	Quinic acid utilization activator	<i>Neurospora crassa</i>	h (A)	a (R)	e (E)
PIP2_YEAST	Peroxisome proliferation regulator	<i>Saccharomyces cerevisiae</i>	h (A)	a (K)	g (T)
STB5_YEAST	Stb5 protein	<i>Saccharomyces cerevisiae</i>	h (L)	a (K)	a (K)
Subfamily 3					
AFLR_EMENI	Sterigmatocystin biosynthesis regulat	<i>Emericella nidulans</i>	c (S)	a (K)	h (V)
AFLR_ASPFL	Aflatoxin biosynthesis regulatory prote	<i>Aspergillus flavus</i>	c (S)	a (K)	h (V)
Individual proteins					
SEF1_YEAST	Suppressor protein Sef1	<i>Saccharomyces cerevisiae</i>	b (H)	a (K)	h (I)
Conservation score			66.6	84.6	63.3

Table A1.22. Helix-loop-helix family (Multi-specific)

The helix-loop-helix family comprises the DNA-binding region of numerous transcription factors that are involved in regulation of cell growth, proliferation, differentiation and apoptosis. Binding is through insertion of long α -helices in the DNA major groove.

Three positions from the probe helix bind the target site. Position 18 presents either a histidine (subfamily 1) or a hydrophobic residue (subfamilies 2 and 3). Proteins with histidine slightly favour guanine (-G...-) while those with hydrophobic residues do not show any preference. Position 22 consists of glutamic acid, which binds cytosine and adenine (-.CA.-) in a complex interaction. Position 26 consists of either arginine (subfamily 1) or a hydrophobic amino acid (subfamilies 2 and 3); in proteins with the first, guanine is favoured (-..G-) while with the latter, no preference is apparent^{60,61}. If asparagine is found at this position, it could be expected to recognise adenine.

Mutagenesis studies have shown that point mutations at the above target-contacting positions alter binding specificity⁶². Mutations are observed for paralogues within species; for example, both 1hloA and MYOG_HUMAN share the same core target sequence -nCAN- but each favours distinct bases at the variable positions. Orthologues from different organisms are well-conserved. Proteins form both homo- and heterodimers which increases the variety of target sequences that can be bound and the regulatory responses that are obtained on DNA-binding.

The lack of different members of the family have been implicated in a wide range of diseases including abnormalities in germ cell development, Williams syndrome and prostate cancer⁶³⁻⁶⁵. However, it is yet unclear whether the disorders can also be caused by point mutations at the target-contacting positions in the proteins.

Table A1.22

Helix-loop-helix family (Multi-specific)										
Representative structure:	1hlo chain A									
Probe α -helix:	Residues 11-39									
Target-contacting positions:	18* (-n....-) 22* (-.CA.-) 26* (-...n-)									
Target sequence:	-nCAN-									
SWISS-PROT/ PDB code	Protein name	Species	Residue positions (numbered with respect to 1hlo chain A)							
			14	15	17	18*	19	22*	25	26*
Subfamily 1										
MAX_BRARE	Max protein	<i>Brachydanio rerio</i>	a (K)	a (R)	b (H)	b (H)	d (N)	e (E)	a (R)	a (R)
MXI1_BRARE	Max interacting protein 1	<i>Brachydanio rerio</i>	h (Y)	a (R)	g (T)	b (H)	d (N)	e (E)	a (R)	a (R)
SRE1_CRIGR	Sterol regulatory element protein-1	<i>Cricetulus griseus</i>	a (K)	a (R)	h (A)	b (H)	d (N)	e (E)	h (Y)	a (R)
SRE2_CRIGR	Sterol regulatory element protein-1	<i>Cricetulus griseus</i>	a (R)	a (R)	g (T)	b (H)	d (N)	e (E)	h (Y)	a (R)
1hloA	Max protein	<i>Homo sapiens</i>	a (K)	a (R)	b (H)	b (H)	d (N)	e (E)	a (R)	a (R)
1an4A	Usf protein	<i>Homo sapiens</i>	a (R)	a (R)	d (Q)	b (H)	d (N)	e (E)	a (R)	a (R)
MAD_HUMAN	Myogenin	<i>Homo sapiens</i>	c (S)	a (R)	g (T)	b (H)	d (N)	e (E)	a (R)	a (R)
USF1_HUMAN	Mad protein (Max dimerizer)	<i>Homo sapiens</i>	a (R)	a (R)	d (Q)	b (H)	d (N)	e (E)	a (R)	a (R)
MXI1_HUMAN	Myogenin Myf-4	<i>Homo sapiens</i>	d (N)	a (R)	g (T)	b (H)	d (N)	e (E)	a (R)	a (R)
USF2_HUMAN	Max interacting protein 1	<i>Homo sapiens</i>	a (R)	a (R)	d (Q)	b (H)	d (N)	e (E)	a (R)	a (R)
1an2A	Max protein	<i>Mus musculus</i>	a (K)	a (R)	b (H)	b (H)	d (N)	e (E)	a (R)	a (R)
USF_STRPU	Upstream stimulatory factor 2	<i>S. purpuratus</i>	a (R)	a (R)	g (T)	b (H)	d (N)	e (E)	a (R)	a (R)
Subfamily 2										
MYF5_BOVIN	Myogenic factor Myf-5	<i>Bos taurus</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MYOD_CAEBR	Myoblast determination protein 1	<i>Caenorhabditis briggsae</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MYOD_CAEEL	Myoblast determination protein 1	<i>Caenorhabditis elegans</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MYOG_COTJA	Myogenic factor 2	<i>Coturnix japonica</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MYOD_DROME	Sterol regulatory element protein-2	<i>Drosophila melanogaster</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MYOG_CHICK	Helix-loop-helix protein dellilah	<i>Gallus gallus</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MYF6_CHICK	Myogenic-determination protein	<i>Gallus gallus</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MYOG_HUMAN	Myogenin Myf-6 (muscle-specific)	<i>Homo sapiens</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
1mdyB	Max protein	<i>Mus musculus</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
MF25_XENLA	Upstream stimulatory factor 2	<i>Xenopus laevis</i>	a (R)	a (R)	h (A)	h (A)	g (T)	e (E)	a (R)	h (L)
Subfamily 3										
DEI_DROME	Sterol regulatory element protein-2	<i>Drosophila melanogaster</i>	a (R)	a (R)	g (T)	h (A)	d (N)	e (E)	a (R)	i (M)
NDF1_HUMAN	Upstream stimulatory factor 1	<i>Homo sapiens</i>	a (R)	a (R)	a (K)	h (A)	d (N)	e (E)	a (R)	i (M)
MTH1_MOUSE	Helix-loop-helix protein 1 hen1	<i>Mus musculus</i>	a (R)	a (R)	h (A)	h (A)	d (N)	e (E)	a (R)	i (M)
MTH2_MOUSE	Neurogenic differentiation factor 1	<i>Mus musculus</i>	a (R)	a (R)	e (E)	h (A)	d (N)	e (E)	a (R)	i (M)
NDF1_XENLA	Helix-loop-helix protein Math-1	<i>Xenopus laevis</i>	a (R)	a (R)	a (K)	h (A)	d (N)	e (E)	a (R)	i (M)
Individual proteins										
HEN1_HUMAN	Upstream stimulatory factor 1 Usf-1	<i>Homo sapiens</i>	h (Y)	a (R)	h (A)	b (H)	h (A)	e (E)	a (R)	h (V)
Conservation score			75.9	100	62.8	71.6	74.5	100	90.7	64.1

References

- 1 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. *Nucleic Acids Res* **29**(13), 2860-74.
- 2 Rice, P. A. (1997). Making DNA do a U-turn: IHF and related proteins. *Curr Opin Struct Biol* **7**(1), 86-93.
- 3 Liang, H., Mao, X., Olejniczak, E. T., Nettesheim, D. G., Yu, L., Meadows, R. P., Thompson, C. B. & Fesik, S. W. (1994). Solution structure of the ets domain of Fli-1 when bound to DNA. *Nat Struct Biol* **1**(12), 871-5.
- 4 Kodandapani, R., Pio, F., Ni, C. Z., Piccialli, G., Klemsz, M., McKercher, S., Maki, R. A. & Ely, K. R. (1996). A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS- domain-DNA complex. *Nature* **380**(6573), 456-60.
- 5 McKercher, S. R., Torbett, B. E., Anderson, K. L., Henkel, G. W., Vestal, D. J., Baribault, H., Klemsz, M., Feeney, A. J., Wu, G. E., Paige, C. J. & Maki, R. A. (1996). Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *Embo J* **15**(20), 5647-58.
- 6 Gruss, P. & Walther, C. (1992). Pax in development. *Cell* **69**(5), 719-22.
- 7 Xu, W., Rould, M. A., Jun, S., Desplan, C. & Pabo, C. O. (1995). Crystal structure of a paired domain-DNA complex at 2.5 Å resolution reveals structural basis for Pax developmental mutations. *Cell* **80**(4), 639-50.
- 8 Balling, R., Deutsch, U. & Gruss, P. (1988). Undulated, a mutation affecting the development of the mouse skeleton, has a point mutation in the paired box of Pax 1. *Cell* **55**(3), 531-5.
- 9 Baldwin, C. T., Hoth, C. F., Amos, J. A., da-Silva, E. O. & Milunsky, A. (1992). An exonic mutation in the HuP2 paired domain gene causes Waardenburg's syndrome. *Nature* **355**(6361), 637-8.
- 10 Hoth, C. F., Milunsky, A., Lipsky, N., Sheffer, R., Clarren, S. K. & Baldwin, C. T. (1993). Mutations in the paired domain of the human PAX3 gene cause Klein- Waardenburg syndrome (WS-III) as well as Waardenburg syndrome type I (WS-I). *Am J Hum Genet* **52**(3), 455-62.

- 11 Tassabehji, M., Read, A. P., Newton, V. E., Harris, R., Balling, R., Gruss, P. & Strachan, T. (1992). Waardenburg's syndrome patients have mutations in the human homologue of the Pax-3 paired box gene. *Nature* **355**(6361), 635-6.
- 12 Tassabehji, M., Read, A. P., Newton, V. E., Patton, M., Gruss, P., Harris, R. & Strachan, T. (1993). Mutations in the PAX3 gene causing Waardenburg syndrome type 1 and type 2. *Nat Genet* **3**(1), 26-30.
- 13 Hanson, I. M., Fletcher, J. M., Jordan, T., Brown, A., Taylor, D., Adams, R. J., Punnett, H. H. & van Heyningen, V. (1994). Mutations at the PAX6 locus are found in heterogeneous anterior segment malformations including Peters' anomaly. *Nat Genet* **6**(2), 168-73.
- 14 Lawson, C. L. & Carey, J. (1993). Tandem binding in crystals of a trp repressor/operator half-site complex. *Nature* **366**(6451), 178-82.
- 15 Joachimiak, A., Haran, T. E. & Sigler, P. B. (1994). Mutagenesis supports water mediated recognition in the trp repressor- operator system. *Embo J* **13**(2), 367-72.
- 16 Gunes, C., Staacke, D., von Wilcken-Bergmann, B. & Muller-Hill, B. (1995). The possible roles of residues 79 and 80 of the Trp repressor from Escherichia coli K-12 in trp operator recognition. *Mol Gen Genet* **246**(2), 180-95.
- 17 Gunes, C. & Muller-Hill, B. (1996). Mutants in position 69 of the Trp repressor of Escherichia coli K12 with altered DNA-binding specificity. *Mol Gen Genet* **251**(3), 338-46.
- 18 Cho, Y., Gorina, S., Jeffrey, P. D. & Pavletich, N. P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **265**(5170), 346-55.
- 19 Martin, A. C., Facchiano, A. M., Cuff, A. L., Hernandez-Boussard, T., Olivier, M., Hainaut, P. & Thornton, J. M. (2002). Integrating mutation data and structural analysis of the TP53 tumor- suppressor protein. *Hum Mutat* **19**(2), 149-64.
- 20 Suckow, M., Kisters-Woike, B. & Hollenberg, C. P. (1999). A novel feature of DNA recognition: a mutant Gcn4p bZip peptide with dual DNA binding specificities dependent of half-site spacing. *J Mol Biol* **286**(4), 983-7.
- 21 Angel, P., Imagawa, M., Chiu, R., Stein, B., Imbra, R. J., Rahmsdorf, H. J., Jonat, C., Herrlich, P. & Karin, M. (1987). Phorbol ester-inducible genes

- contain a common cis element recognized by a TPA-modulated trans-acting factor. *Cell* **49**(6), 729-39.
- 22 Nakabeppu, Y. & Nathans, D. (1989). The basic region of Fos mediates specific DNA binding. *Embo J* **8**(12), 3833-41.
- 23 Konig, P. & Richmond, T. J. (1993). The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. *J Mol Biol* **233**(1), 139-54.
- 24 Koldin, B., Suckow, M., Seydel, A., von Wilcken-Bergmann, B. & Muller-Hill, B. (1995). A comparison of the different DNA binding specificities of the bZip proteins C/EBP and GCN4. *Nucleic Acids Res* **23**(20), 4162-9.
- 25 Matsumoto, T., Nakashima, N., Takase, K., Hirochika, H. & Mizuno, H. (1997). A mutation study of the DNA binding domain of human papillomavirus type11 E2 protein. *J Biochem (Tokyo)* **121**(1), 138-44.
- 26 Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J. & Dickerson, R. E. (1996). How proteins recognize the TATA box. *J Mol Biol* **261**(2), 239-54.
- 27 Kosa, P. F., Ghosh, G., DeDecker, B. S. & Sigler, P. B. (1997). The 2.1-Å crystal structure of an archaeal preinitiation complex: TATA- box-binding protein/transcription factor (II)B core/TATA-box. *Proc Natl Acad Sci U S A* **94**(12), 6042-7.
- 28 Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993). Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**(6446), 512-20.
- 29 O'Brien, R., DeDecker, B., Fleming, K. G., Sigler, P. B. & Ladbury, J. E. (1998). The effects of salt on the TATA binding protein-DNA interaction from a hyperthermophilic archaeon. *J Mol Biol* **279**(1), 117-25.
- 30 Rodriguez-Esteban, C., Tsukui, T., Yonei, S., Magallon, J., Tamura, K. & Izpisua Belmonte, J. C. (1999). The T-box genes Tbx4 and Tbx5 regulate limb outgrowth and identity. *Nature* **398**(6730), 814-8.
- 31 Smith, J. (1999). T-box genes: what they do and how they do it. *Trends Genet* **15**(4), 154-8.
- 32 Li, Q. Y., Newbury-Ecob, R. A., Terrett, J. A., Wilson, D. I., Curtis, A. R., Yi, C. H., Gebuhr, T., Bullen, P. J., Robson, S. C., Strachan, T., Bonnet, D., Lyonnet, S., Young, I. D., Raeburn, J. A., Buckler, A. J., Law, D. J. & Brook,

- J. D. (1997). Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat Genet* **15**(1), 21-9.
- 33 Basson, C. T., Huang, T., Lin, R. C., Bachinsky, D. R., Weremowicz, S., Vaglio, A., Bruzzone, R., Quadrelli, R., Lerone, M., Romeo, G., Silengo, M., Pereira, A., Krieger, J., Mesquita, S. F., Kamisago, M., Morton, C. C., Pierpont, M. E., Muller, C. W., Seidman, J. G. & Seidman, C. E. (1999). Different TBX5 interactions in heart and limb defined by Holt-Oram syndrome mutations. *Proc Natl Acad Sci U S A* **96**(6), 2919-24.
- 34 Ghosh, G., van Duyne, G., Ghosh, S. & Sigler, P. B. (1995). Structure of NF-kappa B p50 homodimer bound to a kappa B site. *Nature* **373**(6512), 303-10.
- 35 Liu, J., Sodeoka, M., Lane, W. S. & Verdine, G. L. (1994). Evidence for a non-alpha-helical DNA-binding motif in the Rel homology region. *Proc Natl Acad Sci U S A* **91**(3), 908-12.
- 36 Schumacher, M. A., Choi, K. Y., Zalkin, H. & Brennan, R. G. (1994). Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science* **266**(5186), 763-70.
- 37 Arvidson, D. N., Lu, F., Faber, C., Zalkin, H. & Brennan, R. G. (1998). The structure of PurR mutant L54M shows an alternative route to DNA kinking. *Nat Struct Biol* **5**(6), 436-41.
- 38 Glasfeld, A., Koehler, A. N., Schumacher, M. A. & Brennan, R. G. (1999). The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions. *J Mol Biol* **291**(2), 347-61.
- 39 Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* **240**(5), 421-33.
- 40 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 angstroms resolution: a complete picture of the protein-DNA interface. *J Mol Biol* **260**(3), 395-408.
- 41 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. *Nucleic Acids Res* **18**(23), 6853-6.

- 42 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 U S A* **87**(12), 4717-21.
- 43 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**(10), 1153-62.
- 44 Parkinson, G., Gunasekera, A., Vojtechovsky, J., Zhang, X., Kunkel, T. A., Berman, H. & Ebright, R. H. (1996). Aromatic hydrogen bond in sequence-specific protein DNA recognition. *Nat Struct Biol* **3**(10), 837-41.
- 45 Zhang, X. P. & Ebright, R. H. (1990). Substitution of 2 base pairs (1 base pair per DNA half-site) within the *Escherichia coli* lac promoter DNA site for catabolite gene activator protein places the lac promoter in the FNR regulon. *J Biol Chem* **265**(21), 12400-3.
- 46 Grindley, N. D. (1993). Analysis of a nucleoprotein complex: the synaptosome of gamma delta resolvase. *Science* **262**(5134), 738-40.
- 47 Hughes, R. E., Rice, P. A., Steitz, T. A. & Grindley, N. D. (1993). Protein-protein interactions directing resolvase site-specific recombination: a structure-function analysis. *Embo J* **12**(4), 1447-58.
- 48 Kim, C. A. & Berg, J. M. (1995). Serine at position 2 in the DNA recognition helix of a Cys2-His2 zinc finger peptide is not, in general, responsible for base recognition. *J Mol Biol* **252**(1), 1-5.
- 49 Desjarlais & Berg, J. M. (1992). Redesigning the DNA-binding specificity of a zinc finger protein: a data base-guided approach. *Proteins* **13**(3), 272.
- 50 Choo, Y., Sanchez-Garcia, I. & Klug, A. (1994). In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* **372**(6507), 642-5.
- 51 Choo, Y. & Klug, A. (1994). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc Natl Acad Sci U S A* **91**(23), 11163-7.
- 52 Choo, Y. & Klug, A. (1994). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A* **91**(23), 11168-72.

- 53 Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R. & Sigler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **352**(6335), 497-505.
- 54 Gewirth, D. T. & Sigler, P. B. (1995). The basis for half-site specificity explored through a non-cognate steroid receptor-DNA complex. *Nat Struct Biol* **2**(5), 386-94.
- 55 Rastinejad, F., Perlmann, T., Evans, R. M. & Sigler, P. B. (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature* **375**(6528), 203-11.
- 56 Latchman, D. S. (1996). Transcription-factor mutations and disease. *N Engl J Med* **334**(1), 28-33.
- 57 Marmorstein, R., Carey, M., Ptashne, M. & Harrison, S. C. (1992). DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **356**(6368), 408-14.
- 58 Vashee, S., Xu, H., Johnston, S. A. & Kodadek, T. (1993). How do "Zn2 cys6" proteins distinguish between similar upstream activation sites? Comparison of the DNA-binding specificity of the GAL4 protein in vitro and in vivo. *J Biol Chem* **268**(33), 24699-706.
- 59 Zhang, L. & Guarente, L. (1994). The yeast activator HAP1--a GAL4 family member--binds DNA in a directly repeated orientation. *Genes Dev* **8**(17), 2110-9.
- 60 Ferre-D'Amare, A. R., Prendergast, G. C., Ziff, E. B. & Burley, S. K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* **363**(6424), 38-45.
- 61 Ferre-D'Amare, A. R., Pognonec, P., Roeder, R. G. & Burley, S. K. (1994). Structure and function of the b/HLH/Z domain of USF. *Embo J* **13**(1), 180-9.
- 62 Feldmann, T., Alex, R., Suckow, J., Dildrop, R., Kisters-Woike, B. & Muller-Hill, B. (1993). Single exchanges of amino acids in the basic region change the specificity of N-Myc. *Nucleic Acids Res* **21**(22), 5050-8.
- 63 Gallant, P., Shiio, Y., Cheng, P. F., Parkhurst, S. M. & Eisenman, R. N. (1996). Myc and Max homologs in Drosophila. *Science* **274**(5292), 1523-7.
- 64 Meng, X., Lu, X., Li, Z., Green, E. D., Massa, H., Trask, B. J., Morris, C. A. & Keating, M. T. (1998). Complete physical map of the common deletion

region in Williams syndrome and identification and characterization of three novel genes. *Hum Genet* **103**(5), 590-9.

- 65 Prochownik, E. V., Eagle Grove, L., Deubler, D., Zhu, X. L., Stephenson, R. A., Rohr, L. R., Yin, X. & Brothman, A. R. (1998). Commonly occurring loss and mutation of the MXI1 gene in prostate cancer. *Genes Chromosomes Cancer* **22**(4), 295-304.

Appendix 2

Table A2.1. A summary of the base-pairs in the target sequences that are in contact with the protein. The first two columns give the name of the protein family and the representative PDB structure. We provide the length of the target sequences for each family, the number of base-pairs in contact with the protein and the number of base-pairs that are not in contact with the protein. We also give the number of base-pairs that are variable in the target site definition: these are also separated into those in contact with the protein and those that are not. Base-pairs are almost always variable if not contacted by the protein. They are also sometimes variable if they are interacting.

Table A2.2. List of PDB codes in multi-specific families that contain more than one complex structure. The PDB code, chain or domain ID, subfamily (see appendix 1) and bound DNA sequence is given. Structures in the same subfamily bind the same DNA sequence and structures in different subfamilies bind different DNA sequences.

Table A2.3. The PET91 substitution matrix. The single-letter amino acid codes are provided in the column and row headings. Scores for each pairwise mutation is normalized between 0 and 100.

Table A2.1

Family		Total number of base-pairs			Total number of variable base-pairs		
		Target sequence	Contacted	Non- contacted	Target sequence	Contacted	Non- contacted
Pu1 ETS domain	1pueE	5	2	3	2	-	2
Prd paired domain	1pdnC	15	5	10	13	2	10
Trp repressor	1trrA	4	3	1	2	1	1
Loop-sheet-helix	1tsrB	5	3	2	3	2	1
Leucine Zipper	2dgcA	4	3	1	3	2	1
Papillomavirus-1 E2	2bopA	6	3	3	2	-	2
TBP	1ytbA	8	8	-	4	4	-
T-domain	1xbrA	8	4	4	3	-	3
RHR	1nfkA	5	3	2	1	1	-

Table A2.2

PDB code	Chain/ domain	Subfamily	DNA sequence bound
Homeodomain			
1fjl,1hdd *	A, B	12	-TAAT-
1au7,1oct *	A2	15	-AAAT- (-TAAT-)
1au7	A1	individual	-ATAC-
1oct	A1	individual	-ATGC-
1yrn	A	individual	-TGTA-
1yrn	B	16	-CATC-
$\beta\beta\alpha$-zinc finger			
1aay,1zaa	A1,A3,C1,C3	18	-GCG-
"	A2,C2	7	-TGG-
2drp	A1,B1	11	-GAT-
"	A2,B2	17	-AGG-
1ubd	C1	37	-GAC-
"	C2	35	-GGA-
"	C3	41	-AAT-
1mey	C1,F1	26	-GAA-
"	C2,F2	31	-GCA-
"	C3,F3,G	2	-GAG-
Hormone receptor			
2nll	A,B	1	-AGGTCA-
1hcq	A,B,E,F	1	-AGGTCA-
1lat	A,B	1	-AGGTCA-
1glu	A,B	2	-AGAACA-
HLH			
1hlo,1an2,1an4	A,B	1	-.CAC-
1mdy	A,B	2	-ACAG-

