J. Mol. Biol. (1995) **253**, 370–382

JMB

Comprehensive Analysis of Hydrogen Bonds in Regulatory Protein DNA-Complexes: In Search of Common Principles

Yael Mandel-Gutfreund, Ora Schueler and Hanah Margalit*

Department of Molecular Genetics, Hebrew University-Hadassah Medical School, Jerusalem, 91120 Israel

**Corresponding author*

A systematic analysis of hydrogen bonds between regulatory proteins and their DNA targets is presented, based on 28 crystallographically solved complexes. All possible hydrogen bonds were screened and classified into different types: those that involve the amino acid side-chains and DNA base edges and those that involve the backbone atoms of the molecules. For each interaction type, all bonds were characterized and a statistical analysis was performed to reveal significant amino acid-base interdependence. The interactions between the amino acid side-chains and DNA backbone constitute about half of the interactions, but did not show any amino acid-base correlation. Interactions *via* the protein backbone were also observed, predominantly with the DNA backbone. As expected, the most significant pairing preference was demonstrated for interactions between the amino acid side-chains and the DNA base edges. The statistically significant relationships could mostly be explained by the chemical nature of the participants. However, correlations that could not be trivially predicted from the hydrogen bonding potential of the residues were also identified, like the preference of lysine for guanine over adenine, or the preference of glutamic acid for cystosine over adenine. While Lys × G interactions were very frequent and spread over various families, the Glu × C interactions were found mainly in the basic helix-loop-helix family. Further examination of the side-chain-base edge contacts at the atomic level revealed a trend of the amino acids to contact the DNA by their donor atoms, preferably at position W2 in the major groove. In most cases it seems that the interactions are not guided simply by the presence of a required atom in a specific position in the groove, but that the identity of the base possessing this atom is crucial. This may have important implications in molecular design experiments.

Keywords: protein-DNA interaction; hydrogen bond; DNA binding motifs; specific recognition; statistical analysis

Introduction

Proteins control transcription of genes by binding to specific DNA regulatory sites. Many aspects of specific protein-DNA interaction are currently understood based on biochemical, molecular and structural studies. Recognition of a regulatory site by a protein is achieved by structural complementarity between the DNA and the protein and by specific contacts established by electrostatic interactions, hydrogen bonds, and hydrophobic interactions. Structural studies of the proteins and protein-DNA complexes by X-ray crystallography and NMR indicated that specific recognition is achieved by distinct structural motifs. Based on these DNA-binding motifs the proteins can be classified into families (reviewed by Pabo & Sauer, 1992). In several of these families structural complementarity is accomplished by an alpha-helix fitting snugly into the DNA major groove. These include the helix-turn-helix (HTH; reviewed by Harrison & Aggarwal, 1990), the homeodomain (reviewed by Gehring *et al*., 1994), the zinc finger (TFIIIA like zinc fingers and hormone receptors) (reviewed by Schmiedeskamp & Klevit, 1994), and the basic leucine zipper (bzip) and basic helix-loop-helix (bHLH; reviewed by Ellenberger, 1994; Wolberger, 1994). Recently, it was shown that specific recognition can also be obtained by a

Abbreviations used: HLH, helix-turn-helix; bHLH, basic HLH; bzip, basic leucine zipper; RHH, ribbon-helix-helix.

0022–2836/95/420370–13 $12.00/0

beta-sheet structure, defining new potential families that use this secondary structural element for DNA recognition either in the major groove (the ribbon-helix-helix, RHH, reviewed by Raumann *et al*., 1994a), or in the minor groove (in the TBP-TATA complex, Y. Kim *et al*., 1993; J. L. Kim *et al*., 1993).

Hydrogen bonds constitute the majority of the interactions between the DNA and the protein. Most of these bonds involve the protein side-chains and the DNA atoms at the base edges and in the backbone, but interactions that involve the protein backbone are also found. The contacts that involve the DNA backbone are believed to stabilize the complex and to orient the protein against the DNA in a fixed arrangement, assisting in establishment of the side-chain-base edge specific interactions (Pabo & Sauer, 1992). The hydrogen bonds between the protein side-chains and the DNA base edges dictate a direct readout of the DNA target by the protein. These interactions are primarily determined by the hydrogen bonding capability of the atoms in the amino acid side-chains and DNA base edges, as proposed by Seeman *et al*. (1976). In their study Seeman *et al*. (1976) characterized potential positions for hydrogen bonds in the DNA grooves (W1 and W2 in the major groove and S1 and S2 in the minor groove, as illustrated in Figure 1). Possible pairs of amino acid-base capable of hydrogen bonding could then be predicted, based on the nature of chemical groups that occupy these positions at the edges of the four DNA bases and on the type of atoms in the amino acid side-chains (Table 1). While in principle there are many such amino acid-base combinations, an intriguing question is whether there is any preference for a particular amino acid to interact with a certain base or with a certain position in the DNA groove. Identification of such relationships could lead to the definition of a general code for recognition.

Attempts in this direction were made initially by Matthews (1988), who examined three DNA-protein complexes known at that time and concluded that there is ‘‘no code for recognition’’. In a recent review Pabo & Sauer (1992) summarized the pairs of amino acid-base in eight complexes and pointed out frequent interactions, such as hydrogen bonds involving purines, particularly guanine residues. Suzuki (1994) screened 20 complexes and demonstrated in them interactions that are predicted by the chemical rules. In addition, he suggested that volume considerations may also play a role. The most successful attempts to identify significant relationships between amino acid and base were obtained within families of specific binding motifs (Kisters-Woike *et al*., 1991; Klevit, 1991; Desjarlais & Berg, 1993; Suzuki & Yagi, 1994; Choo & Klug, 1994a,b), when the position of the amino acid in the motif was also taken into account.

The continuous improvement in the experimental methods for structure determination has led to a significant increase in the number of solved DNA-protein complexes, enabling a systematic analysis of all the contacts involved. In the current study we took a statistical approach to characterize the hydrogen bonds between protein and DNA in 28 structurally solved complexes of transcription factors with their DNA targets (Table 2). By compiling all hydrogen bonds in these complexes and analyzing them by various criteria, statistically significant relationships between the DNA bases and protein amino acids could be identified. These regard interactions between amino acid side-chains and DNA base edges, as well as interactions that involve the backbone of the molecules. Such relationships were examined either at the residue level, i.e. the DNA bases and the protein amino acids, or at a more general level that considers the atoms involved and the position in the groove. Water-mediated interactions which in certain complexes were shown to

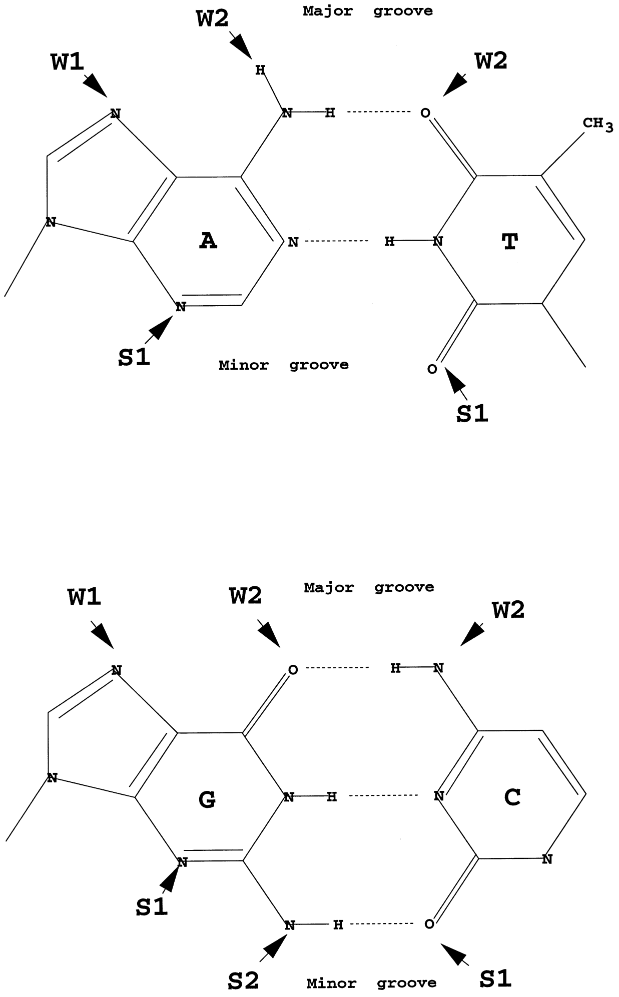

Figure 1. Illustration of the DNA positions in the grooves as determined by Seeman *et al*. (1976). W1 (representing W1 and W1') and W2 (representing W2 and W2') are the outer and central major groove sites, respectively. W1 includes the atom N7 of adenine and guanine, W2 includes O4 of thymidine, O6 of guanine, N6 of adenine and N4 of cytosine. S1 (representing S1 and S1') is the outer minor groove site and includes O2 of thymidine and cytosine and N3 of adenine and guanine. S2 (representing S2 and S2') is the central minor groove site which includes only the atom N2 of guanine.

Table 1. Possible amino acid-base interactions in the major groove (based on hydrogen bonding potential of the atoms)

	A		T	G		C
	W1(a)	W2(d)	W2(a)	W1(a)	W2(a)	W2(d)
Ser(a / d)	+	+	+	+	+	+
Thr(a / d)	+	+	+	+	+	+
Cys(a / d)	+	+	+	+	+	+
Asp(2a)	–	+ +	–	–	–	+ +
Asn(a & d)	+ ⊕	+ ⊕	+	+	+	+
Glu(2a)	–	+ +	–	–	–	+ +
Gln(a & d)	+ ⊕	+ ⊕	+	+	+	+
Lys(d)	+	–	+	+	+	–
Arg(2d)[a]	+ +	–	+ +	+ + ⊕ ⊕	+ + ⊕ ⊕	–
His(2(a / d))	+ +	+ +	+ +	+ +	+ +	+ +
Tyr(a / d)	+	+	+	+	+	+
Trp(d)	+	–	+	+	+	–

W1 and W2 are the specific major groove potential recognition positions as classified by Seeman *et al*. (1976) and illustrated in Figure 1. The hydrogen bonding potential of these positions and of the amino acid side-chains is noted: a, hydrogen bond acceptor; d, hydrogen bond donor. Possible contacts are marked with a plus sign (+), the number of plus signs in a cell represents the number of possible contacts that the participants can make. ⊕ marks interactions that may occur simultaneously.

[a] The atom NE of arginine was not taken into account since it was not found to participate in any of the arginine's interactions with the DNA base edges, probably due to conformational hindrance.

play a role in recognition, have not been included in this study. An extensive analysis was performed for the hydrogen bonds between the DNA base edges and amino acid side-chains. Most of the statistically significant relationships could be explained by the chemical characteristics of the partners in the hydrogen bond. However, significant relationships beyond what is expected by the nature of the chemical groups could also be identified. Although no simple rule could be delineated from the identified correlations, several common principles could be concluded. Some of these characteristics were general and based on the whole database and some could be attributed to specific families.

Results

Screening the 28 crystallographically solved protein-DNA complexes for hydrogen bonds according to the criteria described in Methods has revealed a total of 636 such interactions. Table 3 summarizes all the hydrogen bonds by the pairs of amino acid-base involved.

Analysis of distinct types of DNA-protein contacts

The 636 hydrogen bonds were grouped according to the origin of the atoms participating in the interactions: (1) Protein backbone and DNA backbone (*Pbb* × *Dbb*). (2) Protein backbone and DNA base edge (*Pbb* × *Dbe*). (3) Protein side-chain and DNA backbone (*Psc* × *Dbb*), (4) Protein side-chain and DNA base edge (*Psc* × *Dbe*). The distribution of the interactions in these four categories is shown in Figure 2. As can be seen, the specific interactions between the amino acid side-chains and the DNA base edges constitute only about 30% of the hydrogen bonds. About half of the interactions involve the protein side-chains and the DNA backbone. Interactions that involve atoms in the protein backbone were found predominantly with the DNA backbone rather than with the base edges. The scarcity of *Pbb* × *Dbe* contacts (nine only) is probably due to conformational restrictions that prevent their formation. Most of the interactions with the DNA backbone involve the phosphodiester oxygens (372 out of 439). The interactions with the sugar atoms usually involve atoms that are relatively distant (3.4 Å) and frequently result from bifurcated hydrogen bonds that involve also the phosphate.

By generating a table of pairs of amino acid-base for each type of interaction (similar to Table 3) it is possible to characterize the participants and to look for significant relationships. Tendencies for a particular interaction type can be revealed by looking at the distributions of bases and amino acids that appear in the row and column totals of each table. By comparing such a distribution with the distribution of amino acids and bases in all other interaction types using a χ^2 test, significant preferences of the residues for a certain interaction type can be detected. Interdependence between protein amino acids and DNA bases can be revealed by applying a χ^2 test to each of the tables. In this test the relative abundances of the different amino acids and DNA bases participating in each type of interaction are inherently taken into account, since the expected value in each cell is based on the relative frequencies of the participants in the data. Analysis of individual cells within the tables may point to pairs of amino acid-base that contribute to the statistical significance. Pairs that appear significantly higher than expected are believed to be favorable, while statistically significant infrequent interactions are probably unfavorable. Furthermore, the source of

Table 2. List of DNA-protein complexes used in the analysis

DNA binding motif	Complex	Reference	PDB file
HTH	λ repressor/O_L1	Beamer & Pabo (1992)	1lmb
	434 repressor/O_R1	Aggarwal *et al*. (1988)	2orl
	434 repressor/O_R2	Shimon & Harrison (1993)	1rpe
	434 repressor/O_R3	Rodgers & Harrison (1993)	1per
	434 Cro/O_R1	Mondragon & Harrison (1991)	3cro
	CAP/lac operon (*E. coli*)	Schultz *et al*. (1991)	1cgp
	Trp repressor/operator	Otwinowski *et al*. (1988)	1tro
Homeodomain	POU/Oct-1[b]	Klemm *et al*. (1994)	[a]
	Matα2/operator	Wolberger *et al*. (1991)	[a]
	Engrailed/DNA	Kissinger *et al*. (1990)	1hdd
Zinc finger	TTKDBD/DNA	Fairall *et al*. (1993)	[a]
	zif268/DNA	Pavletich & Pabo (1991)	1zaa
	Gli/DNA	Pavletich & Pabo (1993)	[a]
Hormone receptors	Estrogen receptor/DNA	Schwabe *et al*. (1993)	[a]
	Glucocorticoid receptor/ DNA	Luisi *et al*. (1991)	1glu
Leucine zippers	GCN4/ATF/CREB	König & Richmond (1993)	1dgc
	GCN4/AP-1	Ellenberger *et al*. (1992)	1ysa
bHLH	Max/DNA	Ferré-D'Amaré *et al*. (1993)	[a]
	USF/DNA	Ferré-D'Amaré *et al*. (1994)	[a]
	E47/E-box	Ellenberger *et al*. (1994)	[a]
	MyoD/DNA	Ma *et al*. (1994)	[a]
β-Sheet	TBP/TATA box (CYC1)	Y. Kim *et al*. (1993)	[a]
	TBP/TATA box (Adenovirus)	J. L. Kim *et al*. (1993)	[a]
RHH	Met J repressor/operator	Somers & Phillips (1992)	1cma
	Arc repressor/operator	Raumann *et al*. (1994b)	[a]
Other α-helix	E2/E2-Bs	Hegde *et al*. (1992)	2bop
	Hnf-3/DNA	Clark *et al*. (1993)	[a]
	GaL4/DNA	Marmorstein *et al*. (1992)	1d66

[a] The coordinate files were kindly provided by the author.

[b] The POU/Oct-1 complex is composed of two separate domains: the POU specific domain was classified with the HTH family, while the POU homeodomain was classified with the homeodomain family.

each significant interaction may also be examined, i.e. whether this relationship is attributed to a specific binding motif or is found in various families, indicating a more general correlation.

Protein backbone with DNA backbone (Pbb × Dbb)

Examination of the distribution of the bases in the column totals indicates that guanine is less favorable in this type of interaction (Table 4). Overall, in *Pbb* × *Dbb* interactions the pyrimidines are somewhat more frequent than the purines (opposite to the trend in Table 3), even though not statistically significant. Comparison of the amino acid totals in this table with their distribution in all other interaction types shows that Gly, Ala, and Val participate predominantly in backbone-backbone interactions. The ''preference'' of these amino acids for *Pbb* × *Dbb* interactions is obvious from their chemical nature. They lack side-chain atoms with hydrogen donors or acceptors, and therefore can participate in hydrogen bonds only through their backbone atoms. Their preference over other hydrophobic amino acids may be due to their smaller size. Arginine and lysine also occur frequently, but not as frequent as in interactions that involve their positively charged side-chains. The participation of arginine in this type of interaction is significantly lower than in the other types of interactions. Methionine and phenylalanine that are scarce in protein-DNA interactions are found only in this type of interaction. All the six interactions of methionine are in the Arc repressor-operator complex. Table 4 examines whether there is any dependence between the amino acids and DNA bases that are contacted by their backbone atoms. A χ^2 test reveals statistical significance ($p = 0.005$) and examination of the table suggests some preference for Ser × G, Thr × A, Asn × G and Gln × T. The Gln × T interactions are all attributed to the HTH family, Asn × G are found only in proteins that bind the DNA by a β-sheet while Ser × G are found in both. 25% of the *Pbb* × *Dbb* interactions are coupled to a specific interaction involving the side-chain of the same amino acid. Interestingly, in this type of interaction serine always makes bifurcated hydrogen bonds with one phosphodiester oxygen, while glutamine makes bridging contacts with two consecutive phosphodiester oxygens on the same strand of the DNA.

Table 3. Observed hydrogen bonds between DNA bases and amino acids

	A	T	G	C	Total
Gly	4	5	0	2	11
Ala	4	2	0	2	8
Val	2	2	2	0	6
Leu	0	0	0	0	0
Ile	0	0	0	0	0
Ser	11	13	19	9	52
Thr	16	10	5	7	38
Cys	1	0	0	2	3
Met	2	3	0	1	6
Pro	0	0	0	0	0
Asp	2	0	0	5	7
Asn	28	11	10	5	54
Glu	1	0	0	9	10
Gln	24	16	11	1	52
Lys	11	24	32	19	86
Arg	51	72	90	52	265
His	4	5	4	2	15
Phe	0	2	0	1	3
Tyr	3	5	7	4	19
Trp	0	0	1	0	1
Total	164	170	181	121	636

Protein backbone with DNA base edge (*Pbb* × *Dbe*)

Only nine such interactions are present in the database (Figure 2). Four of these interactions are found in the Matα2-operator complex and involve the small amino acids Ala and Gly, each interacting with adenine and thymidine in the minor groove. The other five interactions are Lys × C and are found in the complexes of λ repressor-operator and Gal4-DNA. The long side-chain of Lys is flexible and therefore may assist the protein backbone in approaching the base edge of the relatively small cytosine. The fact that only these three amino acids form *Pbb* × *Dbe* interactions supports the conjecture that their scarcity is due to conformational restrictions.

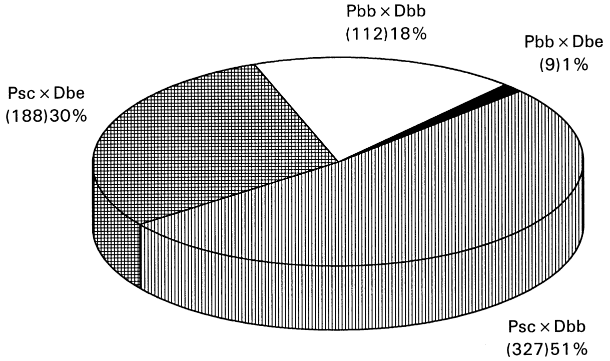

Figure 2. Distribution of hydrogen bonds in protein-DNA complexes classified by the source of atoms involved: (1) *Pbb* × *Dbb*: protein backbone with DNA backbone. (2) *Pbb* × *Dbe*: protein backbone with DNA base edge. (3) *Psc* × *Dbb*: amino acid side-chains with DNA backbone. (4) *Psc* × *Dbe*: amino acid side-chain with DNA base edge.

Table 4. Protein backbone × DNA backbone interactions (*Pbb* × *Dbb*)

	A	T	G	C	*total*
Gly	3	4	0	2	**9↑**
Ala	3	1	0	2	**6↑**
Val	2	2	2	0	**6↑**
Ser	2	1	**6↑**	2	11
Thr	**6↑**	0	1	1	8
Cys	0	0	0	2	2
Met	2	3	0	1	**6↑**
Asn	1	1	**3↑**	0	5
Gln	0	**7↑**	1	0	8
Lys	2	8	2	5	17
Arg	6	10	3	7	**26↓**
Phe	0	2	0	1	**3↑**
Tyr	2	1	1	1	5
total	29	40	**19↓**	24	112

The χ^2 value of the table is 62.42 (p = 0.005). Cells that contribute significantly to the dependence beween protein amino acids and DNA bases ($p \leqslant 0.01$) are marked with bold lines. Boldfaced numbers in the column / row totals represent DNA bases / protein amino acids, respectively, whose frequencies deviate significantly from the expectancy (for calculation of the expected values see Methods). The direction of the arrow (↑,↓) indicates whether the number of interactions was higher or lower than expected.

Protein side-chain with DNA backbone (*Psc* × *Dbb*)

These constitute the majority of interactions (Table 5). Their importance in establishing the complex was observed in many solved complexes and discussed earlier by Pabo & Sauer (1992). As in the other interactions involving the DNA backbone (*Pbb* × *Dbb*) the involvement of guanine in *Psc* × *Dbb* interactions is lower in comparison to its involvement in interactions that involve the base edges. Among

Table 5. Protein side-chain × DNA backbone interactions (*Psc* × *Dbb*)

	A	T	G	C	Total
Ser	8	10	8	6	32
Thr	6	8	4	5	23
Asn	10	2	5	1	**18↓**
Gln	10	4	4	1	19
Lys	8	15	6	8	37
Arg	37	49	40	45	**171↑**
His	4	4	3	2	**13↑**
Tyr	1	4	6	2	13
Trp	0	0	1	0	1
Total	84	96	**77↓**	70	327

Boldfaced numbers in the colum / row totals represent DNA bases / protein amino acids, respectively, whose frequencies deviate significantly from the expectancy (for calculation of the expected values see Methods). The direction of the arrow (↑,↓) indicates whether the number of interactions was higher or lower than expected.

Table 6. Protein side-chain × DNA base edge interactions (*Psc* × *Dbe*)

	A	T	G	C	*total*
Ser	1	2	5	1	9
Thr	2 (2)	1 (1)	0	1	4 (3)
Cys	1	0	0	0	1
Asp	2	NR	NR	**5**↑	**7**↑
Asn	**13** (4)↑	4 (4)	**2**↓	4	**23** (8)↑
Glu	1	NR	NR	**9**↑	**10**↑
Gln	**14**↑	5	**6**↓	0	**25**↑
Lys	**1**↓	1	**24**↑	NR(1)	26 (1)
Arg	**4** (4)↓	8 (5)	**45** (2)↑	0	57 (11)
His	0	1	1	0	2
Tyr	0	0	0	1	1
total	39(10)	**22** (10)↓	**83** (2)↑	**21** (1)↓	165 (23)

The number of interactions in the major groove is shown. NR (not relevant) denotes chemically impossible pairings. Numbers in brackets represent contacts in the DNA minor groove. The χ^2 value of the Table is 139.9 ($p < 10^{-4}$). Cells that contribute significantly to the dependence between protein amino acids and DNA bases ($p \leqslant 0.01$) are marked: dark gray cells denote values that are significantly higher than expected by the frequencies of the participants, while light gray cells indicate values that are significantly low. Boldfaced numbers in the column/row totals are as explained in Tables 4 and 5.

the amino acids, the aromatic ones participate frequently in this type of contact, probably by stacking interactions. Most remarkable is histidine: 13 out of the 15 interactions involving histidine are between its side-chain and the DNA backbone. Arginine is also a major contributor to these interactions, being attracted by its positive charge to the negative phospodiester oxygens of the DNA backbone. Is there any preference of a particular amino acid to interact with the backbone of a certain base? Such a correlation may support the idea of indirect readout through the DNA backbone. However, a χ^2 test on the data in Table 5 showed no dependence between the amino acids and DNA bases. This shows that the different amino acids do not distinguish between the backbone atoms of the four bases, and hence, that indirect readout through the DNA backbone is not guided by simple amino acid-base preferences.

Protein side-chain with DNA base edge (*Psc* × *Dbe*)

These interactions are believed to play a key role in recognition (Table 6). Guanine participates favorably in this type of interaction due to its suitability for interactions with the positively charged amino acids through their pair of hydrogen bond donors. Thymidine and cytosine appear to be less favorable in interactions that involve the protein side-chains and DNA base edges. The low frequency of pyrimidines in *Psc* × *Dbe* interactions can be explained by the presence of only one atom in the major groove that is capable of hydrogen bonding. Examination of the row totals of Table 6 and comparison with the distribution of the amino acids in all other interaction types show that Asp, Glu, Asn and Gln participate in these interactions in highly significant frequencies. Lys and Arg are very frequent in side-chain base edge contacts (although not to a statistical significance).

As predicted, the highest dependence between DNA bases and amino acids was observed for this type of interaction (Table 6). Comparison between Tables 3 and 6 suggests that most of the outstanding cells in Table 3 are due to the specific side-chain-base edge interactions. The statistically significant cells in Table 6 were determined by a χ^2 test that compared the observed frequency of a cell to the expected based on the total frequencies of the bases and amino acids in the table. In general, these statistically significant interactions were consistent with the capability of the participants for hydrogen bonding. However, it was interesting to see whether there were correlations that could not be explained by such considerations *per se*. For this we calculated the expected number of interactions for each pair of amino acid-base based on their hydrogen bonding capability (summarized in Table 1 and exemplified in Methods). By using a binomial test to compare between the observed and expected frequencies, distinct interactions beyond the chemical expectancy could be detected. These interactions are summarized below and the values of statistical significance by the binomial test are reported. Each such interaction was further examined to see whether it was distributed in different families or was a distinct feature of a particular family (Table 7).

(1) Glu × C: In general, interactions that involve the negatively charged Glu are unfavorable due to

Table 7. Distribution of the significant amino acid-base interactions in the different protein families

	HTH	Homeo-domain	Zinc finger	Hormone receptor	bzip	bHLH	β-Sheet	RHH	Others	Total
Glu × C	1	0	0	1	0	7	0	0	0	9
Asp × C	0	0	5	0	0	0	0	0	0	5
Asn × A	0	5	4	0	0	0	4	1	3	17
Gln × A	10	0	0	0	0	0	0	4	0	14
Lys × G	6	0	4	4	0	0	0	3	7	24
Arg × G	9	3	17	4	3	5	0	4	2	47
Others	19	8	13	2	4	5	7	9	5	72
Total	45	16	43	11	7	17	11	21	17	188

Table 8. Interactions between specific atoms in the amino acid side-chains and DNA major and minor groove positions

		W1(a)	W2(a)	W2(d)	S1
Ser	OH	3 A:1/G:2	5 T:2/G:3	1 A:1	
Thr	OH	2 A:2	1 T:1	1 C:1	3 A:2/T:1
Asp	OD			7 A:2/C:5	
Asn	ND2	8 (5) A:2/G:1	5 T:4/G:1		8 A:4/T:4
	OD1			10 (5) A:1/C:4	
Glu	OE			10 A:1/C:9	
Gln	NE2	9 (6) A:1/G:2	9 T:5/G:4		
	OE1			7 (6) A:1	
Lys	NZ	10 A:1/G:9	16 T:1/G:15		1 C:1
Arg	NH	29 (12) A:4/G:13	28 (12) T:8/G:8		11 A:4/T:5/G:2

The first column summarizes the interactions of the amino acids with the outer major groove position (W1) which is occupied exclusively by a hydrogen bond acceptor (a). The second and third columns list the interactions with the inner major groove position (W2), which is occupied by a hydrogen bond acceptor in thymidine and guanine (second column) and by a donor (d) atom in adenine and cytosine (third column). The fourth column lists the interactions with the outer minor groove position (S1). The cells contain information about the total numbers of each interaction (numbers in brackets represent the number of bridging contacts out of the total), and the involvement of the different DNA bases in single bonds.

the negative charge of the DNA. However, when they do occur they involve predominantly cytosine (*via* the N4 atom). This result was found to be statistically significant even when the chemical nature of the participants was taken into account ($p = 0.01$). The only other atom which could interact with Glu in the major groove is the N6 of adenine, but these interactions are rare ($p = 0.01$). The preferred binding to cytosine may be due to its relatively positive charge, attributed to its one donor only (in comparison to adenine with both donor and acceptor atoms in the major groove; Hunter, 1993). The Glu × C interactions are found mainly in the bHLH family (Table 7).

(2) Asp × C: A preference for C as above was observed, however, not to a statistical significance beyond chemical expectancy. These interactions are found only in the zinc fingers.

(3) Asn × A, Gln × A: These interactions are very frequent and appear as highly significant in Table 6, but their frequency is compatible with what is expected from the chemical nature of the participants. Asn and Gln contain donor and acceptor atoms that complement the acceptor and donor of adenine, and could either form single bonds or bridging contacts. While Asn × A interactions were found in various families, the majority of Gln × A hydrogen bonds are attributed to the HTH family (Table 7). As Gln and Asn contain both donor and acceptor atoms they have a potential to interact with all the four bases. However, in some cases the number of such hydrogen bonds was low, below that expected by Table 1 ($p(\mathrm{Asn} \times \mathrm{G}) = 0.04$; $p(\mathrm{Gln} \times \mathrm{C}) = 0.04$).

(4) Lys × G, Arg × G: Guanine contains two acceptors in its major groove that can form a pair of hydrogen bonds with the two donors of Arg and one hydrogen bond with Lys, and indeed these interactions are very frequent. However, their frequency highly exceeds what is expected by the chemical considerations ($p = 0.03$). The scarcity of hydrogen bonds between the side-chains of Arg and Lys and the base edges of adenine and thymidine cannot just relate to the fact that these bases have one acceptor only. These low frequencies are statistically significant when the chemical nature of atoms is taken into account ($p(\mathrm{Arg} \times \mathrm{A}) = 0.03$; $p(\mathrm{Lys} \times \mathrm{A}) = 0.0055$; $p(\mathrm{Lys} \times \mathrm{T}) = 0.0055$). The preference for guanine over adenine and thymidine by the positively charged amino acids may be due to the relatively negative environment of this base, caused by the two acceptors. As can be seen in Table 7, the Lys × G and Arg × G interactions are found in different families and seem to be two of the major interactions that determine specific recognition.

Analysis at the atomic level

The organization of our database of hydrogen bonds enables exploration of the interactions at a more explicit level regarding the atoms involved. By looking at the atomic level it is possible to describe preferred atoms both in the DNA and protein, preferred positions in the DNA grooves, and preferred interactions (Table 8). The majority of interactions are in the major groove. Only 23 of the side-chain-base edge interactions involve atoms in the minor groove, and always in position S1. Most of these interactions are attributed to the homeodomains and to proteins that bind the DNA by a β-sheet structure. Examination of the major groove interactions reveals several interesting trends: (1) 100 interactions are with atoms in the W2 position in comparison to 61 contacts in W1. The predominance of W2 could be explained by the fact that all four bases contain atoms with hydrogen bonding capability in that position, in comparison to W1, which is occupied by such atoms only in the bases G and A (Table 1). But even if we look at acceptor positions only (thus, having equal probabilities for contacts in W1 and W2), more interactions are found in W2, though not statistically significant. For this calculation we considered single hydrogen bonds only, and found that 52 acceptor atoms in W2 in comparison to 38 acceptor atoms in position W1 participate in single hydrogen bonds. (2) Amino acids that contain both donor and acceptor atoms participate predominantly as donors. This again could be tested on single hydrogen bonds. Ser and Thr with their hydroxyl group and Asn and Gln with both donor and acceptor atoms participate in single bonds 31 times as donors and only eight times as acceptors. (3) In principle, it may be possible to identify correlations at the atomic level indicating that an interacting amino acid may not require a specific base but rather a specific atom at a particular position in the groove. However, Table 6 already suggested that at least for Glu, Asp, Lys, and Arg such correlations are ruled out, because significant base preferences were observed. By analyzing the interacting atoms (Table 8) we examined this issue systematically and also noted whether the trends that we found are general or specific to particular families.

(1) Asp and Glu were discussed above. Their side-chains contain acceptor atoms that can interact only with W2 but they interact preferably with C rather than with A (reasoned possibly by the relatively positive environment of C).

(2) Lys: Out of 27 interactions of the side-chain of Lys with the DNA base edges we found only one interaction with A in W1 and one with T in W2. All its other interactions in the major groove are with G, 14 of which evolve from bifurcated hydrogen bonds to W1 and W2 and most of the rest (8/10) are with position W2. The interactions in W1 and W2 are spread across various families of binding motifs, and there is no correlation between the position in the groove and interactions of a particular family.

(3) Arg: Arg × G is found in 12 bridging contacts, but its single bonds show a different preference for G in W1 and W2. While out of its single hydrogen bonds in W1 four are with A and 13 are with G, its 16 single hydrogen bonds in W2 distribute equally between G and T. Thus, G is favored by Arg in position W1 but not in W2. Interestingly, in contrast to the other amino acids whose single hydrogen bonds occur favorably with atoms in W2, Arg does not have any preference to any position in the major groove (17 single hydrogen bonds in W1 and 16 in W2). The bridging contacts involving Arg were found predominantly in the zinc finger family. As for the single bonds, there was no clear tendency for position W1 or W2 in any of the families.

(4) Asn: Despite the prediction that Asn is suitable for interactions with A because of its donor and acceptor atoms that fit the acceptor and donor atoms of A in W1 and W2, respectively, involvement of Asn in bridging contacts was found only five times. However, out of 13 interactions with A in the major groove, ten are accounted to these bridging contacts. Thus, when Asn does interact with A it clearly has a preference for utilizing both donor and acceptor atoms. The single interactions did not show any strong preference to any position in the groove or to any of the Asn atoms, indicating that Asn interacts either as a donor or as an acceptor with no preference. These interactions are spread across the database in different families.

(5) Gln: Gln appears in six bridging contacts with A (four in the HTH family and two in the RHH family). In single bonds Gln shows a clear preference to interact as a donor (12 out of its 13 single interactions), and most of these interactions are in position W2. 21 out of 25 interactions in which Gln is involved are attributed to the HTH family.

(6) Ser and Thr: A clear preference to interact as donors was observed (11 out of 13 interactions in the major groove), however, no clear preference was detected for any of the positions (five interactions in W1 and six in W2 when the amino acids donate a hydrogen atom). Ser interactions were observed predominantly in the zinc finger and also in the HTH families, while Thr interactions occurred in the RHH and HTH families.

Discussion

The current analysis has focused on characterization of the hydrogen bonds in protein-DNA complexes. It is based on an extensive database of 28 complexes including a total of 636 hydrogen bonds and contains precise information for each of the interactions, enabling a systematic analysis of the bonds by various aspects. The results that are reported here concentrate on characterization of the residues that participate in the bonds. In particular, we were interested in identifying base-amino acid relationships that may point to common principles that are used in specific recognition. Even though the term ‘‘recognition code’’ was used before regarding these relationships (Matthews, 1988; Pabo & Sauer, 1992), it was never anticipated to be a precise code like the genetic code or the code by which an mRNA is transcribed on its DNA template. Early evidence suggested that there is no one to one code by which an amino acid recognizes a DNA base (Matthews, 1988), but preferences were predicted based on theoretical considerations (Seeman *et al*., 1976) and observed in smaller sets of crystallographically solved complexes (Pabo & Sauer, 1992; Suzuki, 1994). The current analysis, based on an extended database, enables a precise statistical evaluation of these preferences that is crucial for assessing the significance of the conclusions. It reinforces the previously identified correlations for side-chain-base edge interactions, reveals unexpected preferences (like the preference of Glu for C over A), and identifies additional characteristics of these interactions at the atomic level which were not considered before. In addition, hydrogen bonds involving the backbone atoms of the molecules were also identified and characterized.

We found an interdependence between amino acids and bases in two interaction types: those involving the backbone atoms of the two molecules and those between the amino acid side-chains and DNA base edges. Interactions between the protein backbone and DNA base edges were very rare and appeared only in two protein families, but the ‘‘complementary’’ interactions, between the amino acid side-chains and DNA backbone, constitute about half of all interactions and were found across the whole spectrum of binding motifs. These interactions probably stabilize the complexes and aid in establishing the specific interactions between the side-chains and base edges. No amino acid-base dependence was detected here, suggesting that indirect readout of a DNA target by a protein through interactions with the DNA backbone is not achieved by simple preference of certain amino acids for the backbone of specific bases. Conceivably, in the cases where specific recognition is achieved through such contacts, as proposed by the *trp* repressor/operator cocrystal (Otwinowski *et al*., 1988), the backbone is specified by a particular structure that is sequence-dependent. Since the structure at a certain position may be influenced by neighboring nucleotides, more information can be gained by taking into account the local context of a DNA base. Such an analysis is presently being carried out by us.

The most significant relationships were detected for interactions between the DNA base edges and amino acid side-chains. For this type of hydrogen bonds the statistical analysis was performed in two stages. We first compared the observed frequencies to those expected by statistical considerations and secondly checked whether the statistically significant interactions comply only to what is expected by the hydrogen bonding potential of the participants. We found, as observed before on smaller data sets (Pabo & Sauer, 1992; Suzuki, 1994), that guanine participates frequently in this type of interaction, and that the number of its hydrogen bonds with arginine and lysine exceeds the expectancy based both on statistical considerations and on the chemical nature of the participants. The persistence of the Arg-G, Lys-G correlations with the increase in the database and the fact that these interactions are found in almost all binding motifs may indicate the important role of these amino acid-nucleotide partners in specific protein-DNA interaction. The same is true for Asn × A hydrogen bonds which are found throughout various families. The Asp × C and Glu × C interactions, although not so frequent, are found high above expectancy. However, they are mostly restricted to two families (Asp × C in the zinc fingers and Glu × C in the bHLH family). One possibility is that the less frequent interactions (presumably less favorable energetically) are used to distinguish between different binding motifs and different DNA targets. This applies, for example, to Asp × C, Glu × C, and His × G. The His × G interaction which was found to play a major role in the complex of zif268 with its DNA target (Choo & Klug, 1994b), was not found in any other protein-DNA complex (Table 6). Thus, on the one hand His × G was shown to play a significant role in specific recognition and on the other hand it was found to be quite rare in protein-DNA complexes, suggesting that it has a very distinct function.

In analyzing the base edge-side-chain interactions it was interesting to look also beyond the level of the residues *per se*, and to examine the atoms involved and the positions in the DNA grooves. The analysis at the atom level has revealed some general characteristics of protein-DNA interactions. First it was noticed that the amino acids that contain both donor and acceptor atoms participate predominantly as donors (31 donors in comparison to eight acceptors). These include serine and threonine, which can either donate or accept a hydrogen atom by their hydroxyl group, and aspargine and glutamine each possessing both donor and acceptor atoms. Indeed, as predicted by Seeman *et al*. (1976) there are twice as many positions for donors to interact with the DNA than for acceptors (Table 1). However, the ratio of 31:8 exceeds this expectation tremendously. This may be due to the overall negative environment of the DNA and the relatively positive properties of the donor atoms. We also found a tendency for the amino acid side-chains to contact position W2 in the DNA major groove, suggesting that W2 is favored over W1.

An intriguing conjecture is that a protein can bind similarly to various sequences because it recognizes atoms with the same hydrogen bonding capability in identical positions of the DNA grooves. For example, if an acceptor atom is required in position W1, it is possible that it does not matter to the protein whether it comes from an adenine or a guanine. From our results we can identify the interactions where the atom type determines the specificity and those where the identity of the residue is important (Table 8). Interactions where the residue identity is important are best exemplified by the results for lysine (with its one donor atom). In principle it could be expected that in some cases lysine would prefer to bind to position W1 and then interact with A or G (which both possess the N7 atom in this position) and in other cases with position W2 and then interact with G (O6) or T (O4). However, we found only two hydrogen bonds of Lys with A and T, while all its other 24 hydrogen bonds in the major groove were with G. The discrimination against T and A may be due to steric clashes with the methyl and amino groups, respectively, or to a preference of Lys for bifurcated hydrogen bonds with acceptor atoms in W1 and W2 (present only in G). Indeed we found that 14 of the 24 hydrogen bonds appeared in seven bifurcated bonds, suggesting that this may be the explanation for the discrimination against T and A. A similar phenomenon was noticed in a recent experimental study by Choo & Klug (1994b). In their selections for compatible binding sites for the zif268 second zinc finger from libraries of randomized DNA triplets, they found a high discrimination against adenine in the second position of the binding triplet when the interacting amino acid was His, favoring exclusively the His × G interaction. However, in the crystal of zif268 with its DNA target, His was modeled to interact with the N7 atom of G in position W1 (Pavletich & Pabo, 1991), suggesting that adenine can also be a suitable partner. In their paper Choo & Klug (1994b) suggested that the discrimination against adenine could be due to a steric interference with its amino group. Alternatively, if the bond is not solely with W1 as suggested by the crystal, they proposed that a preference for a bifurcated bond may discriminate against adenine. Reinforcement of these conclusions can also be obtained from the interpretation of substitution experiments (Takeda *et al*., 1989; Smith & Johnson, 1994) in view of the solved crystal structures (Mondragon & Harrison, 1991; Wolberger *et al*., 1991). Apparently, for most of the interactions the presence of a specific base and not only an appropriate atom is crucial. These conclusions have important implications in molecular design procedures. Since His × G interactions appear only once in our database it is hard to draw general conclusions but rather to restrict the implications to DNA sequences that may interact with zif268-like sequences. As for Lys, it can be definitely stated that in the design of specific protein binding sequences, certain positions that need to interact with Lys should preferably contain G. On the other hand, Table 8 shows that in some instances where the hydrogen bond involves a donor atom of the protein interacting with an acceptor in position W2, the bases G and T may appear interchangeably. This is seen in single bonds in W2 involving Ser, Gln, and Arg. Thus, design of DNA sequences where it is known that the protein needs to bind position W2 by any of these amino acids may contain either G or T.

From Tables 6 and 8 it is clear that the preferences for hydrogen bonds between amino acids and bases are mostly based on chemical considerations. Amino acids that are capable of double hydrogen bonds will contact preferably a base with complementary chemical groups in W1 and W2 (e.g. Arg × G, Asn × A, or Gln × A). Amino acids may prefer one base over another because the environment of that base is more favorable (Asp and Glu prefer C over A), or because it enables bifurcated bonds (preference of Lys for G). Still, a simple general code that applies to all cases and can be used in prediction cannot be defined. Recognition patterns for a specific binding motif family (distinct from a general code) look more tractable (Kisters-Woike *et al*., 1991; Desjarlais & Berg, 1993; Choo & Klug, 1994b; Jamieson *et al*., 1994). However, even in the case of specific recognition patterns, attempts to use theoretical considerations that encompass the size of the residues and their chemical characteristics did not lead to satisfactory predictions (Suzuki & Yagi, 1994), as is evident from the discrepancy between such predictions for TFIIIA-like zinc fingers (Suzuki & Yagi, 1994) and the recognition pattern that was defined experimentally for zif268 contacts with DNA targets (Rebar & Pabo, 1993; Choo & Klug, 1994a,b; Jamieson *et al*., 1994). The experimentally defined recognition pattern for zif268 and its derivatives (obtained by substitutions) fulfills the chemical considerations discussed above. However, it is not clear why certain residues and not others are allowed in certain positions. Also, that code, as the authors cautiously state (Choo & Klug, 1994b), may not apply to all TFIIIA-like zinc fingers, but to those that resemble zif268. Indeed, the defined patterns are not present in all sequences in a database of 385 zinc fingers (Rosenfeld & Margalit, 1993), suggesting that other amino acid-base pairing may also be used, even within this family. Probably, other factors such as short and long range interactions with other amino acids or the conformation of the DNA, also play a significant role in specific recognition but cannot be taken into account at this state of knowledge. In summary, while amino acid-base preferences could be identified, it seems that based on the currently available data, a simple recognition code that can be generally applied cannot be delineated.

Methods

Data organization

Our data set contained 28 coordinate files of crystallographically solved protein-DNA complexes. Only complexes of regulatory proteins with their DNA targets were included (Table 2). Most of the proteins belong to well defined families of DNA binding motifs. The coordinate files were either extracted from the Protein Data Bank, PDB (Bernstein *et al*., 1977), or obtained directly from the authors, as listed in Table 2. In the few cases where data were available for the same protein binding to different sites, all the complexes were included in the analysis. These include the bacteriophage 434 repressor with its three different operators O_R1,O_R2,O_R3 (Aggarwal *et al*., 1988; Shimon & Harrison, 1993; Rodgers & Harrison, 1993), and the two complexes of the GCN4 protein with different target sites (König & Richmond, 1993; Ellenberger *et al*., 1992). Similarly, complexes that involve homologous proteins from different organisms with their target sites, like the TBP (Y. Kim *et al*., 1993; J. L. Kim *et al*., 1993) and the bacteriophages' repressors (Aggarwal *et al*., 1988; Beamer & Pabo, 1992) were all included. Equivalent interactions in complexes of oligomers which bind a symmetrical or partially symmetrical DNA site were considered only once.

Determination of hydrogen bonds

Hydrogen bonds with a maximum distance of 3.5 Å between acceptor and donor atoms were included. Ippolito *et al*. (1990) determined 3.4 Å as the maximum distance for a hydrogen bond, but since in the current analysis some of the complexes were of lower resolution, we used 3.5 Å as a threshold. Determination of the amino acid side-chain and DNA base edge atoms that can participate in hydrogen bonds was based on Ippolito *et al*. (1990) and Seeman *et al*. (1976), respectively. For characterization of the positions in the grooves where interactions occur we used the terminology of Seeman *et al*. (1976), demonstrated in Figure 1. In addition we examined the participation of the backbone atoms in hydrogen bonds: the donor N–H of the protein backbone, and the acceptors, O=of the protein backbone, O^- of the phosphodiester and –O– of the deoxyribose DNA backbone.

Database of hydrogen bonds

All hydrogen bonds in the complexes were extracted and characterized. For each bond the following information was recorded. (1) The distance between acceptor and donor atoms. (2) The pair of amino acid-base participants (amino acids are designated in three letter code and DNA bases in one letter code). (3) Specific atoms in the amino acid and DNA base that participate in the hydrogen bond. (4) Affiliation to a binding motif family.

Analysis of hydrogen bonds

The sort of information recorded for each hydrogen bond enables statistical analyses of the data either at the level of the residues involved (amino acids and bases), or at a more precise level regarding the atoms. It is possible to analyze the database of hydrogen bonds as a whole, or to refer to sections of it that are specified by any of the recorded parameters. We divided the interactions into four types according to the source of atoms involved. Interactions of the protein side-chains with the DNA base edges ($Psc \times Dbe$) or with the DNA backbone ($Psc \times Dbb$), and interactions that involve the protein backbone and the DNA base edges ($Pbb \times Dbe$) or the protein backbone and the DNA backbone ($Pbb \times Dbb$). For each interaction type we generated a sub-database of hydrogen bonds and searched for amino acid-base correlations that deviate from random. $Psc \times Dbe$ interactions were analyzed not only at the residue level but also at the level of the atoms involved, by considering the position in the DNA groove. Since the protein family affiliation was recorded for each bond, it was possible to determine if an identified amino acid-base correlation is general or specific to a certain family.

Statistical analysis

In order to search for significant correlations the data was arranged in two-way contingency tables, rows and columns representing amino acids and bases, respectively. A Pearson χ^2 test was performed to determine whether the row and column classifications were independent. Significant χ^2 values ($p \leqslant 0.01$) indicated a correlation between the DNA bases and the protein amino acids participating in the hydrogen bonds. In tables that showed such interdependence, the χ^2 statistic (which is calculated as the sum of χ^2 over all cells) also enables us to identify which cells contributed to the statistical significance (for each cell $p_{\mathrm{df}=1} \leqslant 0.01$), and to detect important amino acid-base relationships.

As in most of the tables the data were too sparse, a standard χ^2 test to obtain the asymptotic p-value was not adequate and the exact permutational distribution of the Pearson χ^2 statistic had to be calculated (Metha *et al*., 1990). The exact p-value was computed on an IBM compatible PC computer using the StatXact package (Metha & Patel, 1991). When an exact p-value was calculated for the whole table, the p-value ($df = 1$) for the individual cells was calculated accordingly. For this, a 2×2 contingency table was constructed for each cell, where X_{11} contained the number of observations in the examined cell, and the other three cells contained the complements for this value when the row, column and the whole table totals are considered. In Results the p-value was denoted by the letter p independent on its method of calculation (either in a standard fashion or by the exact method).

To detect a preference of a certain amino acid or a DNA base to participate in each of the different interaction types ($Psc \times Dbe$, $Psc \times Dbb$, $Pbb \times Dbe$, $Pbb \times Dbb$), we compared the frequencies in the row and column totals in each table to the total frequencies in all other interaction types. A χ^2 test for goodness of fit was carried out in order to detect significant differences. Again, for comparisons that differed significantly, it was possible to specify the amino acids or DNA bases that contributed to the significance by the individual χ^2 values ($df = 1$).

The hydrogen bonding potential of the amino acid side-chain and DNA base edge atoms enable only certain $Psc \times Dbe$ interactions. In order to find out whether the distribution of the different amino acid-base interactions was guided solely by the hydrogen bonding potential of the participants we have calculated the expected frequencies of the various pairs based on the chemical nature of the participants, summarized in Table 1. For example, the relative potentials of lysine to interact with A:T:G:C is 1:1:2:0, respectively, thus the theoretical probability of lysine to interact with A, T, G, or C is 0.25:0.25:0.5:0, respectively. By multiplying the theoretical probabilities by the total number of interactions involving lysine (26) the expected numbers of Lys × A, Lys × T, Lys × G, Lys × C can be obtained. Differences between expected and observed frequencies in each row of the $Psc \times Dbe$ table (Table 6), considering only contacts to the major groove, were evaluated using a binomial test. Comparisons which generated a p value $\leqslant 0.05$ in the binomial test were considered as favorable or unfavorable interactions, beyond what is expected by the chemical properties of the participants.

Acknowledgements

We are grateful to Professor Norman Grover for statistical consultation and for the use of the statistical package StatXact. This work was supported by The Israel Science Foundation.

References

Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science*, **242**, 899–907.

Beamer, L. J. & Pabo, C. O. (1992). Refined 1.8 Å crystal structure of the λ repressor-operator complex. *J. Mol. Biol.* **227**, 177–196.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Choo, Y. & Klug, A. (1994a). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl Acad. Sci. USA*, **91**, 11163–11167.

Choo, Y. & Klug, A. (1994b). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl Acad. Sci. USA*, **91**, 11168–11172.

Clark, K. L., Halay, E. D., Lai E. & Burley, S. K. (1993). Co-crystal structure of the HNF-3 / *fork head* DNA-recognition motif resembles histone H5. *Nature*, **364**, 412–420.

Desjarlais, J. R. & Berg, J. M. (1993). Use of the zinc finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.

Ellenberger, T. (1994). Getting a grip on DNA recognition: structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. *Curr. Opin. Struct. Biol.* **4**, 12–21.

Ellenberger, T. E., Brandl, C. J., Struhl K. & Harrison S. C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: crystal structure of the protein-DNA complex. *Cell*, **71**, 1223–1237.

Ellenberger, T., Fass, D., Arnaud, M. & Harrison, S. C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Genes Dev.* **8**, 970–980.

Fairall, L., Schwabe, J. W. R., Chapman, L., Finch, J. T. & Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger / DNA recognition. *Nature*, **366**, 483–487.

Ferré-D'Amaré, 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**, 38–45.

Ferré-D'Amaré, 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**, 180–189.

Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G. & Wüthrich, K. (1994). Homeodomain-DNA recognition. *Cell*, **78**, 211–223.

Harrison, S. C. & Aggarwal, A. K. (1990). DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* **59**, 933–969.

Hegde, R. S., Grossman, S. R., Laiminis, L. A. & Sigler, P. B. (1992). Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature*, **359**, 505–512.

Hunter, C. A. (1993). Sequence-dependent DNA structure: the role of base stacking interactions. *J. Mol. Biol.* **230**, 1025–1054.

Ippolito, J. A., Alexander, R. S. & Christianson, D. W. (1990). Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.* **215**, 457–471.

Jamieson, A. C., Kim, S.-H. & Wells, J. A. (1994). *In vitro* selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, **33**, 5689–5695.

Kim, J. L., Nikolov, D. B. & Burley S. K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, **365**, 520–527.

Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993). Crystal structure of a yeast TBP / TATA-box complex. *Nature*, **365**, 512–519.

Kissinger, C. R., Liu B., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell*, **63**, 579–590.

Kisters-Woike, B., Lehming, N., Sartorius, J., von Wilcken-Bergmann, B. & Müller-Hill, B. (1991). A model of the *lac* repressor-operator complex based on physical and genetic data. *Eur. J. Biochem.* **198**, 411–419.

Klemm, J. D., Rould, M. A., Aurora, R., Herr, W. & Pabo, C. O. (1994). Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell*, **77**, 21–32.

Klevit, R. E. (1991). Recognition of DNA by Cys_2 His_2 zinc fingers. *Science*, **253**, 1367, 1393.

König, 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**, 139–154.

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**, 497–505.

Ma, P. C. M., Rould, M. A., Weintraub, H. & Pabo, C. O. (1994). Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell*, **77**, 451–459.

Marmorstein R., Carey, M., Ptashne, M. & Harrison, S. C. (1992). DNA recognition by Gal4: structure of a protein-DNA complex. *Nature*, **356**, 408–414.

Matthews, B. W. (1988). No code for recognition. *Nature*, **335**, 294–295.

Metha, C. & Patel, N. (1991). StatXact statistical software for exact nonparametric inference. Cytel software corporation, Cambridge, MA.

Metha, C., Patel N. & Senchaadhuri, P. (1990). Exact significance testing by the method of control variates. Physica Verlag. *Proc. COMPSTAT*. 141–144.

Mondragon, A. & Harrison, S. C. (1991). The phage 434 Cro / O_R1 complex at 2.5 Å resolution. *J. Mol. Biol.* **219**, 321–334.

Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. & Sigler, P. B. (1988). Crystal structure of *trp* repressor / operator complex at atomic resolution. *Nature*, **335**, 321–329.

Pabo, C. O. & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053–1095.

Pavletich, N. P. & Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*, **252**, 809–817.

Pavletich, N. P. & Pabo, C. O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science*, **261**, 1701–1707.

Raumann, B. E., Brown, B. M. & Sauer, R. T. (1994a). Major groove DNA recognition by β-sheets: the ribbon-helix-helix family of gene regulatory proteins. *Curr. Opin. Struct. Biol.* **4**, 36–43.

Raumann, B. E., Rould, M. A., Pabo, C. O. & Sauer, R. T. (1994b). DNA recognition by β-sheets in the Arc repressor-operator crystal structure. *Nature*, **367**, 754–757.

Rebar, E. J. & Pabo, C. O. (1993). Zinc finger phage: affinity selection of fingers with new DNA-binding specifities. *Science*, **263**, 671–673.

Rodgers, D. W. & Harrison, S. C. (1993). The complex between phage 434 repressor DNA-binding domain and operator site O_R3: structural differences between consensus and non-consensus half-sites. *Structure*, **1**, 227–240.

Rosenfeld, R. & Margalit, H. (1993). Zinc fingers: conserved properties that can distinguish between spurious and actual DNA-binding motifs. *J. Biomol. Struct. Dynam.* **11**, 557–570.

Schmiedeskamp, M. & Klevit, R. E. (1994). Zinc finger diversity. *Curr. Opin. Struct. Biol.* **4**, 28–35.

Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). Crystal structure of a CAP-DNA complex: the DNA is bent by 90°. *Science*, **253**, 1001–1007.

Schwabe, J. W. R., Chapman, L., Finch, J. T. & Rhodes, D. (1993). The crystal structure of the Estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell*, **75**, 567–578.

Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.

Shimon, L. J. W. & Harrison, S. C. (1993). The phage 434 O_R2/R1-69 complex at 2.5 Å resolution. *J. Mol. Biol.* **232**, 826–838.

Smith, D. L. & Johnson, A. D. (1994). Operator-constitutive mutations in a DNA sequence recognized by a yeast homeodomain. *EMBO J.* **13**, 2378–2387.

Somers, W. S. & Phillips, S. E. V. (1992). Crystal structure of the *met* repressor-operator complex at 2.8 Å resolution reveals DNA recognition by β-strands. *Nature*, **359**, 387–393.

Suzuki, M. (1994). A framework for the DNA-protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317–326.

Suzuki, M. & Yagi, N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.

Takeda, Y., Sarai, A. & Rivera, V. M. (1989). Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Natl Acad. Sci. USA*, **86**, 410–443.

Wolberger, C. (1994). b/HLH without the zip. *Struct. Biol.* **1**, 413–416.

Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D. & Pabo, C. O. (1991). Crystal structure of a MAT α2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell*, **67**, 517–528.

Edited by P. E. Wright

(Received 27 March 1995; accepted in revised form 4 August 1995)