J. Mol. Biol. (1998) 277, 1129–1140

JMB

A Role for CH···O Interactions in Protein–DNA Recognition

Yael Mandel-Gutfreund[1], Hanah Margalit[1]*, Robert L. Jernigan[2] and Victor B. Zhurkin[2]*

[1]*Department of Molecular Genetics and Biotechnology The Hebrew University–Hadassah Medical School POB 12272, Jerusalem 91120 Israel*

[2]*Laboratory of Experimental and Computational Biology National Cancer Institute National Institutes of Health Bldg. 12B, Rm. B116 MSC 5677, 12 South Drive Bethesda, MD 20892-5677 USA*

**Corresponding authors*

The concept of CH···O hydrogen bonds has recently gained much interest, with a number of reports indicating the significance of these non-classical hydrogen bonds in stabilizing nucleic acid and protein structures. Here, we analyze the CH···O interactions in the protein–DNA interface, based on 43 crystal structures of protein–DNA complexes. Surprisingly, we find that the number of close intermolecular CH···O contacts involving the thymine methyl group and position C5 of cytosine is comparable to the number of protein–DNA hydrogen bonds involving nitrogen and oxygen atoms as donors and acceptors. A comprehensive analysis of the geometries of these close contacts shows that they are similar to other CH···O interactions found in proteins and small molecules, as well as to classical NH···O hydrogen bonds. Thus, we suggest that C5 of cytosine and C5-Met of thymine form relatively weak CH···O hydrogen bonds with Asp, Asn, Glu, Gln, Ser, and Thr, contributing to the specificity of recognition. Including these interactions, in addition to the classical protein–DNA hydrogen bonds, enables the extraction of simple structural principles for amino acid–base recognition consistent with electrostatic considerations.

Keywords: protein–DNA recognition; hydrogen bonds; CH···O interactions; electrostatics

Introduction

Specific binding of proteins to DNA regulatory elements plays a central role in the control of many cell processes. Specificity is achieved by surface shape complementarity between the protein binding domain and the DNA grooves, as well as by specific contacts involving the amino acid side-chains and the DNA bases, mainly through hydrogen bonds. By definition, hydrogen bonds are formed between two electronegative atoms sharing a proton between them, where one of the participants is the donor and the other is the acceptor of the proton (Pauling, 1960). Usually, nitrogen atoms and oxygen atoms are involved in such interactions. However, there is growing evidence of the formation of weak attractive CH···O hydrogen bonds, first detected in the high resolution crystal structures of organic compounds (Sutor, 1962; Taylor & Kennard, 1982; Desiraju, 1991).

Although for many years little attention has been paid to these non-classical hydrogen bonds in biological macromolecules, recently a number of reports have indicated the significance of these interactions for stabilizing nucleic acid and protein structures (reviewed by Wahl & Sundaralingam, 1997). Wahl *et al*. (1996) discovered a novel U·U base-pair (denoted the Calcutta base-pair) in an RNA hexamer crystal structure that is stabilized by a non-conventional C5-H···O2 bond in addition to a conventional N3-H···O4 hydrogen bond. Interstrand CH···O hydrogen bonds are also presumed to stabilize the intercalated cytosine-rich DNA quadruplex structure (Berger *et al*., 1996). CH···O and CH···N interactions in the major groove involving C5 of cytosine and C5-Met of thymine were postulated to account for the specific recognition between the nucleic acid helices during recombination (Zhurkin *et al*., 1994). The importance of CH···O interactions in proteins is also evident from a recent analysis of 13 high resolution protein crystal structures (Derewenda *et al*., 1995). In their study Derewenda and coworkers showed that many short CH···O interactions in proteins exhibit stereochemical features typical of hydrogen bonds. Bella & Berman (1996) identified two repetitive patterns of CH···O hydrogen bonds in the collagen triple-helix, which are believed to be sig-

0022–2836/98/151129–12 $25.00/0/mb981660

nificant for stabilizing the structure. In the present study we explore the existence and possible role of CH···O hydrogen bonds within protein–DNA complexes.

Recently, we analyzed and characterized the intermolecular hydrogen bonds in a data set of 28 crystal structures of transcription factor–DNA complexes (Mandel-Gutfreund *et al.*, 1995), in an attempt to reveal general principles that determine specific recognition between bases and amino acid residues. In that study only conventional hydrogen bonds were considered between atoms in the DNA grooves and the amino acid side-chains, as initially proposed by Seeman *et al.* (1976). Our results indicated, however, that these hydrogen bonds by themselves are not sufficient to account for the amino acid–base preferences observed in crystal structures, and that electrostatic effects may also play a significant role. Independently, it was suggested that electrostatic interactions are important for discriminating between the bases and for increasing the selectivity of nucleic acid recognition in recombination (Rao & Radding, 1994; Zhurkin *et al.*, 1994). As the next step, here we use an extended data set of 43 crystallographically solved protein–DNA complexes (Table 1), and examine also interactions that involve position C5 in cytosine, C5(Cyt), and the corresponding position in thymine, occupied by a methyl group, C5M(Thy). We consider only the interactions in the major groove, since they are more abundant compared with those in the minor groove (the latter will be analyzed separately). We show that CH groups in the major groove are rather frequently involved in close contacts with protein atoms, mostly oxygen. Stereochemical analysis of the contacts indicates that their features are similar to the weak CH···O hydrogen bonds observed in small molecules and

Table 1. List of DNA-protein complexes used in the analysis

Binding motif	Complex	Reference	NDB file	Resolution (Å)
Enzymes	*Eco*RI endonuclease/DNA	(Kim *et al.*, 1990)	PDE001	2.5
	*Eco*RV endonuclease/DNA	(Kostrewa & Winkler, 1995)	PDE014	2.0
	*Pvu*II endonuclease/DNA	(Cheng *et al.*, 1994)	PDE017	2.6
	Hin recombinase/DNA	(Feng *et al.*, 1994)	PDE009	2.3
Helix	λ Repressor/O_L1	(Beamer & Pabo, 1992)	PDR010	1.8
turn	434 Repressor/O_R1	(Aggarwal *et al.*, 1988)	PDR004	2.5
helix	434 Repressor/O_R2	(Shimon & Harrison, 1993)	PDR011	2.5
	434 Repressor/O_R3	(Rodgers & Harrison, 1993)	PDR015	2.5
	434 Cro/O_R1	(Mondragon & Harrison, 1991)	PDR001	2.5
	CAP/DNA	(Parkinson *et al.*, 1996)	PDR023	2.5
	trp repressor/operator	(Otwinowski *et al.*, 1988)	PDR009	1.9
	trp repressor/operator	(Lawson & Carey, 1993)	PDR013	2.4
Homeo-domain	Pou/Oct-1	(Klemm *et al.*, 1994)	PDT019	3.0
	Paired domain/DNA	(Xu *et al.*, 1995)	PDR018	2.5
	Pax/DNA	(Wilson *et al.*, 1995)	PDE025	2.0
	MATα2/operator	(Wolberger *et al.*, 1991)	PDT005	2.7
	MAT a1-MAT α2/DNA	(Li *et al.*, 1995)	PDT028	2.5
	eve/DNA	(Hirsch & Aggarwal, 1995)	PDT031	2.0
	Engrailed/DNA	(Kissinger *et al.*, 1990)	PDT004	2.8
Zinc	TTK/DNA	(Fairall *et al.*, 1993)	PDT011	2.8
finger	Zif268/DNA	(Elrod-Erickson *et al.*, 1996)	PDT039	1.6
	Gli/DNA	(Pavletich & Pabo, 1993)	PDT008	2.6
Hormone	Estrogen receptor/DNA	(Schwabe *et al.*, 1993)	PDRC03	2.4
receptors	Glucocorticoid receptor/DNA	(Luisi *et al.*, 1991)	PDRC01	2.9
Leucine	GCN4/ATF/CREB	(Konig & Richmond, 1993)	PDT007	3.0
zippers	GCN4/AP-1	(Ellenberger *et al.*, 1997)	PDT002	2.9
	C-Fos-C-Jun/DNA	(Glover & Harrison, 1995)	PDT014	3.05
Basic	Max/DNA	(Ferré-D'Amaré *et al.*, 1993)	PDT023	2.9
helix–loop–	USF/DNA	(Ferré-D'Amaré *et al.*, 1994)	PDT027	3.1
helix	e47/E-box	(Ellenberger *et al.*, 1994)	PDT020	2.8
	MyoD/DNA	(Ma *et al.*, 1994)	PDT016	2.8
β sheet	TBP/TATA box (CYC1)	(Kim *et al.*, 1993)	PDT012	2.5
	TBP/TATA box (Adenovirus)	(Kim & Burley, 1994)	PDT009	1.9
	Human TATA box	(Nikolov *et al.*, 1996)	PDT034	1.9
Ribbon	*met* J repressor/operator	(Somers & Phillips, 1992)	PDR008	2.8
helix–helix	Arc repressor/operator	(Raumann *et al.*, 1994)	PDR012	2.6
Others	p53/DNA	(Cho *et al.*, 1994)	PDR022	2.2
	E2/E2-Bs	(Hegde *et al.*, 1992)	PDV001	1.7
	HNF-3/DNA	(Clark *et al.*, 1993)	PDT013	2.5
	GAL4/DNA	(Marmorstein *et al.*, 1992)	PDT003	2.7
	PPR1-DNA	(Marmorstein & Harrison, 1994)	PDT017	3.2
	PurR/DNA	(Schumacher *et al.*, 1994)	PDR020	2.7
	NF-κb/DNA	(Muller *et al.*, 1995)	PDT022	2.6

Forty-three crystal protein–DNA complexes were analyzed. The coordinate files were extracted from the Nucleic Acid Database, NDB (Berman *et al.*, 1992). For consistency, each complex included in the data set was represented by the minimal unit required for specific recognition, regardless of the number of units crystallized (i.e. in the case of proteins which bind the DNA as homodimers, the dimer was included, while proteins which bind the DNA as monomers were included only once).

in recent studies of protein and nucleic acid structures. It is thereby suggested that these CH···O interactions contribute to the specificity of nucleoprotein recognition. Furthermore, re-evaluation of the amino acid–base preferences in the solved protein–DNA complexes, including CH···O interactions, emphasizes the importance of electrostatic considerations in specific recognition.

Results and Discussion

Frequencies of protein–DNA interactions at the atomic level

A detailed inspection of the protein–DNA interface is crucial for understanding the mode of recognition between the two molecules. Surprisingly, an unexpectedly large number of close contacts is observed between the protein oxygen and the DNA base edge carbon atoms (especially the methyl group of thymine (C5M) and the corresponding position C5 in cytosine). Two such contacts, C5M(Thy)···O and C5(Cyt)···O, are illustrated in Figure 1. Our analysis shows that for the A·T pairs the number of C5M(Thy)···O contacts observed at a distance of $\leqslant$3.0 Å is approximately half the number of the classical N6(Ade)···O hydrogen bonds (Figure 2(a)). At distances $\leqslant$3.5 Å these numbers are comparable, and at longer distances the C5M···O contacts dominate. We presume that at shorter distances these contacts could account for weak CH···O hydrogen bonds, while at longer distances they are probably less specific and may not reflect direct contacts. Yet, the high occurrence of these distant contacts suggests that there is a high probability for protein oxygen to be in the vicinity of the thymine methyl groups due to the exposure of these groups in the major groove.

In the case of G·C pairs, position N4 of cytosine is the most attractive for the protein oxygen at the short distances $\leqslant$3.0 to 3.5 Å (Figure 2(b)). However, at the longer distances $\leqslant$4.0 Å the number of contacts between the protein oxygen atoms and atoms N4 and C5 become comparable. In the case of cytosine these interactions occur almost exclusively with the carbonyl (O=) and carboxyl (O^-) groups of the amino acid side-chains and not with the hydroxyl groups, unlike thymine (where the hydroxyl OH groups are found in approximately 50% of all the cases, see Table 2).

In principle, it is possible to interpret these contacts to be a consequence of the attraction between the oxygen and neighboring DNA positions, thus presuming that the C5···O contacts are secondary rather than primary interactions. In particular, in the case of cytosine it is likely that the oxygen would be attracted to the amino group at position N4. In fact, out of 21 C5(Cyt)···O interactions

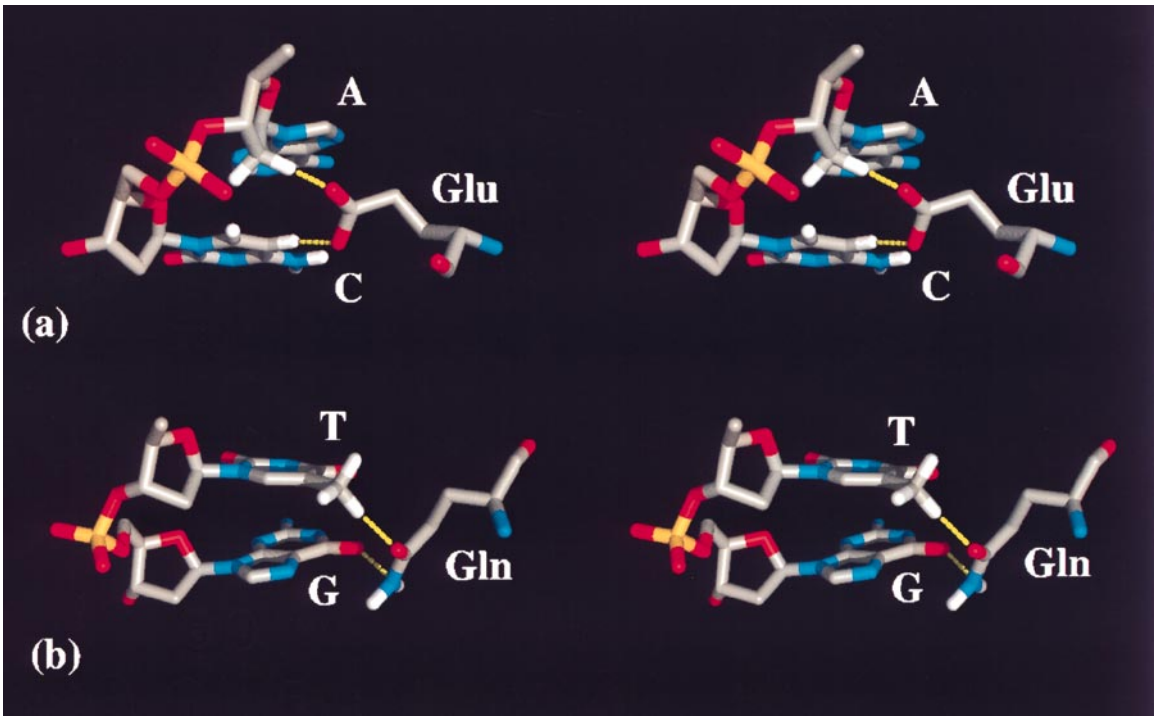

Figure 1. (a) Stereo view of an interaction between the amino acid Glu32 and the dinucleotide AC in the complex between the 434 repressor and the operator OR1 (Aggarwal *et al.*, 1988). The proposed hydrogen bond (2.96 Å) between atoms OE2(Glu) and C5(Cyt) is shown in broken lines. In addition, a close contact (2.7 Å) between the other oxygen atom of Glu and a carbon atom of the sugar ring in the preceding adenosine is also shown, the H···O distance is 1.62 Å. (b) Stereo view of the interaction between Gln29 and the dinucleotide TG in the 434 Cro–OR1 complex (Mondragon & Harrison, 1991). The classical hydrogen bond (3.0 Å) between NH2(Gln) and O6(Gua), and the proposed weak hydrogen bond (2.94 Å) between OE1(Gln) and C5M(Thy) are shown in broken lines.

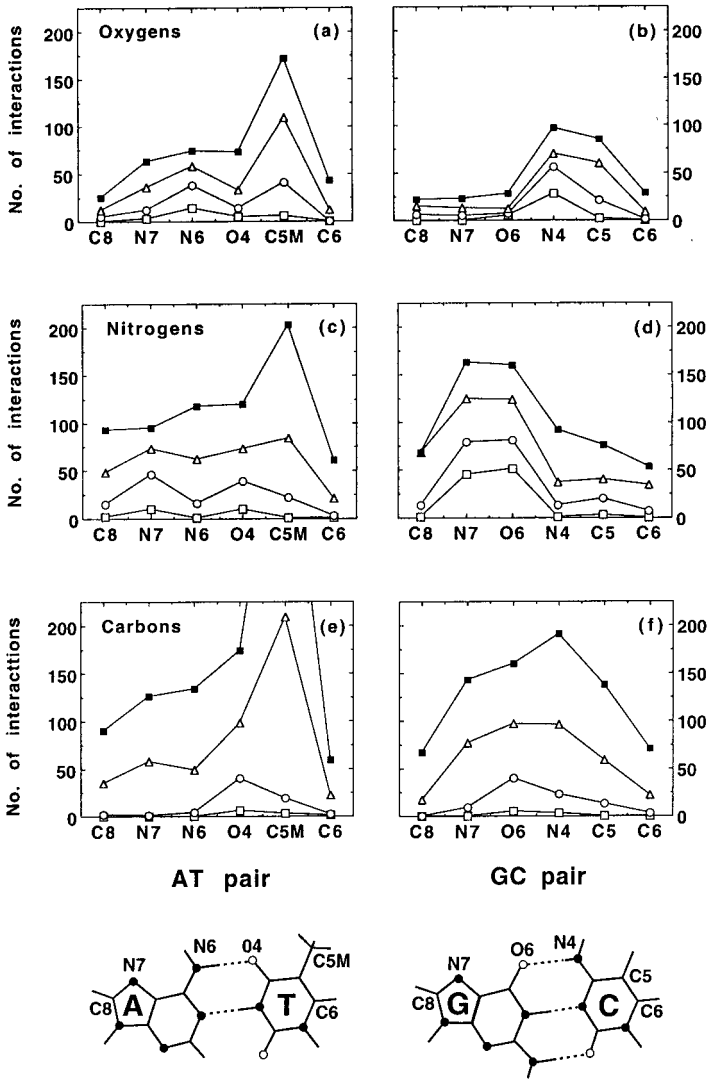

Figure 2. Protein–DNA contacts in the major groove. Number of interactions of the protein oxygens ((a) and (b)), nitrogens ((c) and (d)) and carbons ((e) and (f)) with the DNA atoms exposed in the major groove for A·T pairs (left) and G·C pairs (right). The lines with open squares, circles, triangles and filled squares are for distances $\leqslant$3.0, $\leqslant$3.5, $\leqslant$4.0, and $\leqslant$4.5 Å, respectively.

(Table 2), 13 involve both the C5 and N4 positions while eight interactions are with the C5 alone. In the case of thymine, one should distinguish among the possible interactions with neighboring positions depending on the type of protein oxygen involved. While all oxygen types (OH, O=/O$^-$) can form contacts with neighboring bases, the hydroxyl OH can also participate in hydrogen bonds with position O4 of thymine or with backbone oxygen atoms *via* its hydrogen. Accordingly, out of 18 OH groups that interact with C5M(Thy) (Table 2), 15 are localized so that they can form hydrogen bonds with DNA oxygen atoms, but they do not interact as acceptors with neighboring positions. Out of 23 oxygens of the O=/O$^-$ type (Table 2), three interact with a neighboring base, while 20 are in contact only with the methyl group. Thus, a significant fraction of the interactions between the protein oxygen atoms and the C5(Cyt)/C5M(Thy) do not involve other donor groups and appear *per se* to be attractive.

The number of close contacts ($\leqslant$3.5 Å) observed between the protein nitrogen atoms (representing mostly amino groups) and the DNA major groove positions is consistent with what could be expected based on electrostatic considerations (Figure 2(c) and (d)). Moreover, the interactions with positions N7 and O6 of guanine (Figure 2(d)) are the most abundant, probably due to the overall negative charge of guanine that attracts the positively charged amino acids, as was suggested by Hunter (1993) and emphasized in recent studies (Choo & Klug, 1994; Suzuki, 1994; Lustig & Jernigan, 1995; Mandel-Gutfreund *et al.*, 1995). Interactions that involve the protein nitrogen atoms and the DNA carbon atoms become especially noticeable at longer distances. For example, at distances $\leqslant$4.5 Å the number of interactions C5M(Thy)···N exceeds the number of all other interactions with protein nitrogen atoms (Figure 2(c)). These interactions could be facilitated by the attraction between the protein NH groups and the phosphate oxygen atoms of the DNA backbone or the neighboring O4(Thy) positions in the groove. An alternative explanation could be that this kind of interaction is due to a relatively weak electrostatic attraction between the

Table 2. Angles and distances for NH···O and CH···O contacts[a]

	d, H···O (Å)	D, C/N···O (Å)	ϕ (deg.)	ξ (deg.)	θ_d (deg.)	θ_p (deg.)	No. of contacts
N6(Ade) (3.5 Å)	2.3(±0.4)	3.1(±0.2)	146(±25)	126(±15)	15(±12)	24(±17)	38
(3.2 Å)	2.1(±0.3)	3.0(±0.2)	146(±25)	128(±16)	17(±13)	24(±19)	26
(3.0 Å)	2.0(±0.1)	2.8(±0.1)	152(±21)	130(±11)	15(±13)	24(±21)	14
N4(Cyt) (3.5 Å)	2.2(±0.4)	3.0(±0.3)	145(±28)	118(±29)	20(±19)	35(±21)	56
(3.2 Å)	2.1(±0.3)	2.9(±0.2)	149(±25)	121(±29)	20(±19)	34(±21)	46
(3.0 Å)	1.9(±0.2)	2.8(±0.2)	155(±22)	128(±26)	16(±16)	33(±22)	29
C5(Cyt)	2.4(±0.3)	3.2(±0.1)	136(±21)	118(±)23	23(±20)	43(±23)	21
C5M(Thy) ⟺ O^- & O=	2.4(±0.4)	3.2(±0.3)	145(±25)	110(±29)	39(±29)	46(±26)	23
C5M(Thy) ⟺ OH	2.4(±0.2)	3.3(±0.2)	143(±20)	131(±25)	17(±19)	29(±19)	18
C5M(Thy) ⟺ all O	2.4(±0.3)	3.3(±0.2)	144(±23)	119(±29)	29(±27)	39(±19)	41

Means and standard deviations for angles and distances are given for close contacts ⩽3.5 Å between the heavy atoms. NH···O interactions were classified by N···O distances ⩽3.0, ⩽3.2, ⩽3.5 Å, and means and standard deviations were calculated for each of these data sets. In the case of C5(Cyt) all CH···O interactions were considered as a single set. In the case of thymine the protein carbonyl (O=) and carboxyl (O^-) groups, and the hydroxyl (OH) oxygens were analyzed separately, to make the comparison with cytosine easier, as C5(Cyt) interact with the (O=) and (O^-) groups in 20 cases of 21. ϕ is defined as the angle C5/N4-H···O for cytosine, N6-H···O for adenine and C5M-H···O for thymine. ξ is defined as the angle H···O-C. θ_d and θ_p are the DNA and protein elevation angles, respectively (see Methods).

[a] The hydrogens were built as described in Methods.

deg., degree.

nitrogen and the methyl hydrogen atoms, analogous to that proposed for the amino groups (Sponer & Kypr, 1994; Sponer *et al.*, 1996).

With regard to the contacts involving protein carbon atoms, the C5/C5M···C contacts dominate, probably owing to the hydrophobic effect (Figure 2(e) and (f)). As expected, the number of these interactions is significant only at distances larger than the sum of van der Waals' radii of two carbon atoms, 3.5 Å. In fact, these interactions become overwhelming at the distances ⩽4.5 Å: there are 456 contacts C5M(Thy)···C (data not shown) and 135 contacts C5(Cyt)···C (Figure 2(f)). This resembles the observed preferences in proteins for favorable charged pairs at close approach and hydrophobic pairs at longer distances (Bahar & Jernigan, 1997). A noticeable number (80) of interactions between the protein carbon atoms and the DNA positions occupied by oxygen were also observed at distances less than 3.5 Å (40 contacts with O4(Thy) and 40 with O6(Gua), see Figure 2(e) and (f)). Although these numbers are modest compared with the C···C interactions, they are similar to the number of classical hydrogen bonds, such as N6(Ade)···O(protein). Interestingly, the protein carbon atoms interact very often also with the NH_2 group of cytosine, especially at distances ⩽4.5 Å (191 contacts, Figure 2(f)).

Geometric features of the close CH···O contacts

To further characterize the nature of the C5···O interactions observed between position C5/C5M in the DNA major groove and the protein oxygen atoms, the geometric features of all close contacts (⩽3.5 Å) were studied. The inter-atomic distances and angles of these ''bonds'' were calculated and the mean values were compared with the values obtained for classical NH···O hydrogen bonds in our data set. Since in most of the structures only the heavy atoms were determined, hydrogens were built for both the CH···O and NH···O interactions (see Methods). A summary of the mean values and standard deviations of bond angles and inter-atomic distances obtained for the CH···O and NH···O interactions is given in Table 2.

Contacts via *the C5 group of cytosine*

The H···O distances ~2.45 Å for C5(Cyt)H···O contacts are in good agreement with earlier observations for CH···O interactions in high resolution protein crystal structures (Derewenda *et al.*, 1995) and in small molecules (Taylor & Kennard, 1982). Naturally, these distances are large compared with the corresponding distances in NH···O hydrogen bonds (Table 2). Similarly, the C···O distances obtained directly from the crystallographic data, are also larger than the N···O distances, due to the larger size of carbon (1.75 Å) compared with nitrogen (1.55 Å). These observations suggest that CH groups in the DNA major groove can form hydrogen bonds with the protein oxygen atoms, but, however, these bonds are weaker than conventional hydrogen bonds. This is supported by recent *ab initio* calculations of the imidiazole–water interactions, that estimate the CH···O hydrogen bond energy as 2 to 3 kcal/mol compared with ~7 kcal/mol for the NH···O bond (Ornstein & Zheng, 1997).

The angles calculated for the C5···O contacts in the protein–DNA complexes are also in agreement with the geometries of other CH···O hydrogen bonds reported previously. Although the mean value of the ϕ angle (C5-H···O), 136°, does not fall within the range usually accepted for hydrogen bonds (140° to 180°), this is mainly due to the extremely low values (ϕ < 120°) found in some structures. This may imply that the data set of C5···O interactions is actually composed of two sub-groups, only one of which corresponds to

"real" hydrogen bonds (with $\phi = 140^\circ$ to 180°), as was suggested for the CH···O interactions in proteins (Derewenda *et al.*, 1995). It should be noted that the mean ϕ angle (N4-H···O) observed for the NH···O interactions is also rather low (145° to 155°), although somewhat higher than that for CH···O contacts (Table 2). Some of the NH···O interactions exhibit geometries unacceptable for hydrogen bonds (8 of 38 for adenine and 10 of 56 for cytosine), perhaps due to low resolution structures.

The ξ angle (H···O-C) is quite close to the ideal value for the XH···O hydrogen bonds (120°), implying that the proton is directed toward the oxygen lone electron pair. Furthermore, the angles observed for CH···O interactions are closer to the ideal geometry than those observed for NH···O interactions (Table 2). The DNA elevation angles θ_d are relatively low for both CH···O and NH···O interactions ($\theta_d = 15^\circ$ to 23°), indicating that the protein oxygen atoms lie very close to the base planes. The protein elevation angles θ_p are somewhat higher, 24° to 43°, especially for the CH···O contacts, suggesting that in this type of interaction the proton is positioned away from the plane of the carbonyl/carboxyl group. This deviation from planarity can also be explained by the relatively low resolution of the complexes and the repulsive C···O potentials used in crystallographic refinement (Derewenda *et al.*, 1995).

Contacts involving the thymine methyl group

In the case of thymine, the CH···O interactions were divided into two subgroups, one involving carbonyl (O=) and carboxyl (O^-) oxygen atoms of proteins, and the other involving the hydroxyl groups OH (Table 2). Based on electrostatic considerations, we expect that when interacting with thymine, the hydroxyl groups would approach the methyl group from the center of the base, contacting the two major groove positions simultaneously (C5M and O4), while carbonyl and carboxyl oxygens can approach C5M group only from the edge of the base-pair, avoiding repulsion from the negatively charged oxygen O4. (This distinction was not made in the case of cytosine: first, since the rationale for dividing the contacts into such subgroups does not hold when an NH group is in the neighboring position N4; and second, since in most of the cases the interactions with C5(Cyt) involve carbonyl or carboxyl oxygen atoms only.) Using our method of placing the methyl protons (described in Methods), the H···O and C···O distances are similar to those obtained for cytosine (Table 2). Notice that for carbonyl/carboxyl the standard deviations are always larger than for hydroxyl. This is consistent with the above notion that the hydroxyl can make double contacts with the thymine groups O4 and C5M, and therefore would be preferably located between positions 4 and 5. On the contrary, carbonyl/carboxyl can interact with various DNA groups in addition to C5M(Thy), such as the adjacent base or the sugar-phosphate backbone, and thus its localization is less certain.

The ϕ angles for C5M(Thy)···O contacts are larger than for cytosine, and closer to the ideal geometry with $\phi = 180^\circ$ (especially for carbonyl/carboxyl, see Table 2). In this case the higher ϕ values can be explained by repulsion between the carbonyl/carboxyl group and O4(Thy). On the other hand, the elevation angles θ_d and θ_p are relatively high in the case of thymine, especially for the interactions that involve the carbonyl group. Notice, however, that in this case the protein oxygen does not have to be in the base plane for the energetically favorable interaction with the methyl proton (see Figure 1(b)).

Finally, the analysis of the hydrogen bond parameters indicates that both in CH···O interactions and in the classical NH···O hydrogen bonds the values are very widely dispersed and do not always agree with the ideal hydrogen bond parameters (Table 2). Nevertheless, although our sample is rather limited, the results for CH···O interactions in both cytosine and thymine do resemble observations made on low weight compounds having CH···O hydrogen bonds (Taylor & Kennard, 1982), as well as those on high resolution protein structures (Derewenda *et al.*, 1995; Bella & Berman, 1996). In the case of thymine this is somewhat surprising, since methyl groups are less polarized than the aromatic carbon atoms, and are rarely found to be involved in CH···O interactions in nucleic acids (Wahl & Sundaralingam, 1997). Thus, in our case it remains debatable whether the C5M(Thy)···O contacts should be described as CH···O hydrogen bonds, or as intermediates between weak hydrogen bonds and strong van der Waals' interactions.

Distribution of C5···O distances

Analysis of the inter-atomic distances in the C5···O DNA–protein contacts provides further insight into the nature of these interactions. To characterize these contacts in the DNA–protein interface, the interatomic distances in the range from 2.6 Å (hydrogen bonds) to 6.0 Å (long range interactions) were studied. Here again the C5···O contacts are compared to N···O and C···C interactions. Figure 3 shows the normalized frequency distribution of distances of pairwise contacts (see Methods). As expected, for short distances (less than the sum of the van der Waals' radii of two carbon atoms) the probability for C5···C contacts is rather low. As the distance increases, these contacts become abundant and represent about 60% of all interactions involving C5. Starting at 4.0 Å the frequency of these contacts reaches a plateau that roughly corresponds to the fraction of protein carbon atoms in the DNA–protein interface. The pattern of the N···O interactions is in a sense a mirror image of the C···C profile and reflects the strong attraction between the two groups. Up to a

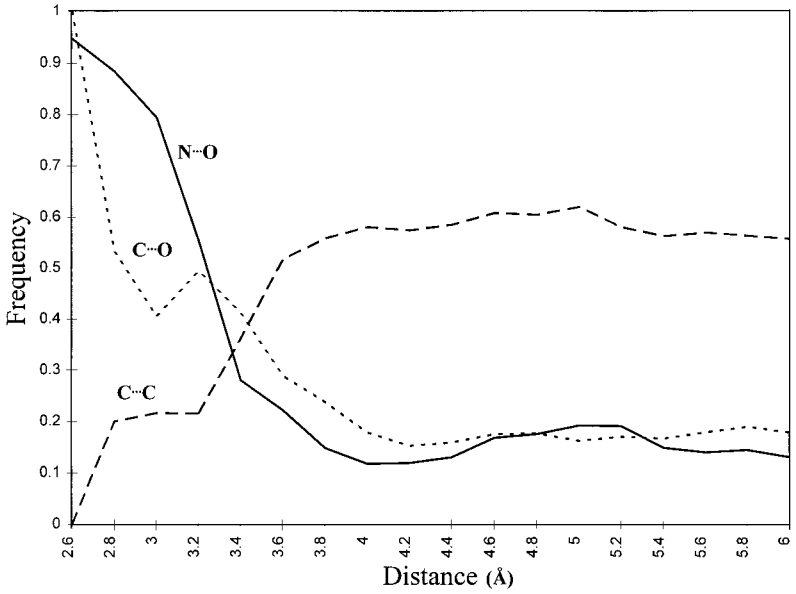

Figure 3. Normalized frequency distributions of distances for pairwise contacts (N···O, C···C, C···O). Dotted line is for N···O pairs (N4(Cyt)/N6(Ade) and protein oxygens); broken line is for C···C pairs (C5(Cyt)/C5M(Thy) and protein carbon atoms); continuous line is for C···O pairs (C5(Cyt)/C5M(Thy) and protein oxygen atoms). Notice that the C···O curve has a profile that is intermediate between those for ''classical'' hydrogen bonds N···O and ''non-specific'' interactions C···C.

distance of 3.2 Å these interactions dominate, but then decrease with a sharp slope, until beyond the distance of 3.8 Å their fraction remains approximately 20% of all interactions involving NH groups in DNA.

The pattern of the C5···O interactions resembles that of N···O. The slope for the C5···O interactions is very steep and somewhat broken, probably as a consequence of the scarce data. Starting from a distance of 3.2 Å this profile is intermediate between the other two (C5···C and N···O), and at a distance of 4.2 Å it reaches a plateau at approximately the same level as for the N···O interactions. The level of the plateau reflects the fraction of protein oxygen atoms in the DNA–protein interface. Notice that in the interval between 3.2 and 4.2 Å there is a shift of 0.2 Å between the graph for the N···O and that for the C5···O, which corresponds to the greater van der Waals' radius of carbon (1.75 Å) compared with oxygen (1.55 Å). The similarity between the two distribution patterns, for C5···O and N···O interactions, is in accord with the data presented in Table 2, and indicates the close relationship between the two types of contacts. These results are entirely consistent with the similar analysis of CH···O interactions in high resolution protein crystal structures (Derewenda *et al.*, 1995), which demonstrated the intermediate characteristics of these interactions compared to N···O and C···C contacts.

In the current study we have used crystallographically solved protein–DNA complexes of transcription factors and restriction enzymes. These include different DNA binding motif families with various numbers of representatives in each family. In order to verify that our findings are not biased, we repeated the above analysis using a data set that includes only one representative from each binding motif family, and all structures in the ''Others'' category in Table 1. The normalized frequency distributions obtained for this limited data set practically coincide with those presented in Figure 3 (data not shown).

Electrostatic considerations in amino acid–base recognition

In search of general principles for specific amino acid–base recognition, we and others noticed earlier that electrostatic considerations may play a significant role (Choo & Klug, 1994; Suzuki, 1994; Lustig & Jernigan, 1995; Mandel-Gutfreund *et al.*, 1995). By considering the C5···O close contacts ($\leqslant$3.5 Å) as part of the protein–DNA hydrogen bonding network, this conjecture is reinforced. We obtained a total of 391 interactions between the amino acid side-chain tips and the DNA base edges in the major groove (54 of which are CH···O interactions), involving 276 different pairs of amino acid–base.

To link the observed frequencies of the pairs of amino acid–base with simple physico-chemical properties of the participants, the DNA bases and the amino acid residues have been arranged according to their charges on the major groove edge and the side-chain tip, respectively . For this purpose, the bases were ranked by their net charges in the major groove, G < A $\leqslant$ T < C, guanine being the most negative and cytosine the most positive (Tables 3 and 4). The amino acids were divided into four major groups according to their chemical properties (Table 3). These groups were arranged according to the charges on the side-chain tips, $(O^-) < (OH) < (O = /NH) < (N^+)$, see Table 5. Histidine was classified separately because of its pH-dependent properties. Aliphatic amino acids were excluded, although in principle they can form weak CH···O hydrogen bonds with oxygen atoms on the DNA base edges, e.g.

Table 3. Occurrences of pairs of amino acid–DNA base in 43 X-ray complexes

A. Classical hydrogen bonds

aa group \ Base	G	A	T	C
O^-	0	1	0	**17**
OH/SH	6	9	7	4
O = /NH	12	**40**	17	8
N^+	**79**	9	15	0
His	**6**	2	3	0

B. Classical hydrogen bonds and CH···O interactions

aa group \ Base	G	A	T	C
O^-	0	1	6	**21**
OH/SH	6	9	**23**	4
O = /NH	12	**40**	*29*	11
N^+	**79**	9	15	0
His	**6**	2	3	0

A. Pairs of amino acid–base, formed by classical hydrogen bonds as postulated by Seeman *et al.* (1976). Hydrogen bonds with a maximum distance of 3.5 Å between acceptor and donor atoms in the major groove were included. B. Inclusion of CH···O interactions (⩽3.5 Å) in addition to the classical hydrogen bonds. Bases are arranged according to their net charge in the major groove, from the negative (left) to the positive (right), see Table 4. Amino acids are divided into four major groups (aa groups), depending on the acceptor/donor propensity of the side-chain tip and ordered according to charge: (O^-), Asp, Glu; (OH/SH), Ser, Thr, Tyr and Cys; (O = /NH), Asn, Gln; (N^+), Arg, Lys (see Table 5). Histidine was classified separately due to its pH-dependent properties.

Bold and underlined letters indicate the highest numbers in each row, showing the preferred base for each amino acid group. The relatively large number of interactions between the O = /NH group of amino acids and thymine in B is indicated in italics. Notice the obvious diagonals across the Table, indicating the consistency of the amino acid preferences with their electrostatic properties, which is much more emphasized in B where CH···O interactions are included.

O4(Thy). However, this type of contact is beyond the scope of our present study and will be discussed elsewhere.

Arranging the data according to the relative charges of the participants in a 4 × 5 Table highlights the obvious diagonal across the Table (from right top to left bottom), indicating the consistency of the amino acid–base preferences with their electrostatic properties (Table 3B). Interestingly, without the C5-H···O interactions included, this pattern is weaker and no preference is found for the (OH) group of amino acids (Table 3B), suggesting that these weak hydrogen bonds can contribute to the specificity of recognition *via* the C5 groups of pyrimidines. Moreover, the proposed attractive interactions between C5M(Thy) and protein oxygen atoms can explain experimentally observed amino acid–base preferences that otherwise remain unaccounted for (e.g. see Choo & Klug, 1994). In their selection studies Choo & Klug (1994) used sequence variants of the zif268 second zinc finger to screen libraries of all possible DNA triplet binding sites. Close inspection of their results (Figure 1, Choo & Klug, 1994) indicates that in most cases Ser and Thr interact often with thymine, as predicted by our scheme (Table 3B). The relatively large number (29) of pairs between amino acids (O = /NH) and thymine in Table 3B, is likewise consistent with the experimentally observed preferences of Asn and Gln for thymine, in addition to the anticipated preference for adenine (Choo & Klug, 1994). Notice that considerations based only on classical hydrogen bonds fail to account for this (O = /NH)···(Thy) preference (Table 3A). Similar amino acid–base preferences are also obtained for the limited set of data (see above), when only one representative from each binding motif family is used. Overall, Table 3 emphasizes that electrostatic considerations are useful for elucidating general guidelines for amino acid–base recognition.

Table 4. Base edge charges in the major groove

Scale	G	A	T	C
Renugopalakrishnan *et al.* (1971)	−0.54	−0.09	−0.04	0.33
Pearlman & Kim (1990)	−1.01	−0.01	0.28	0.78
Cornell *et al.* (1995)	−0.45	0.07	−0.03	0.39

To calculate the net partial charges in the major groove, the following heavy atoms were considered together with the attached protons: adenine, N7, C5, C6, N6; thymine, O4, C4, C5, C5(Met); guanine, N7, C5, C6, O6; and cytosine, N4, C4, C5. Three sets of charges were used. Notice that the only discrepancy between the three sets of charges occurs for adenine and thymine: the potentials by Cornell *et al.* (1995) give the order A > T (underlined) in contrast to the other potentials.

Table 5. Amino acid tip charges

Scale	(OH/SH)	(O = /NH)
Momany *et al.* (1975)	(−0.14)–(−0.17)	0.03–0.04
Cornell *et al.* (1995)	(−0.16)–(−0.27)	(−0.01)–0.04

Two sets of charges were used; both predict the more negative charges on the amino acid tips in the (OH/SH) group. The group (OH/SH) contains Ser, Thr, Cys and Tyr; tips include the O, S and H atoms. The group (O = /NH) contains Asn and Gln; tips include the amino groups, carbonyl oxygens and the adjacent carbons (C^γ in Asn and C^δ in Gln).

Conclusions

The concept of CH···O hydrogen bonds has been debated for many years, since the conjecture of Pauling (1960), who proposed that the difference in the boiling point between trifluoroacetyl chloride and acetyl chloride is due to a hydrogen bond involving a methyl group in the latter (reviewed by Desiraju, 1991). However, in recent years evidence has accumulated that these bonds may have significant implications in various biochemical processes (reviewed by Wahl & Sundaralingam, 1997). In this paper, based on 43 crystal structures of protein–DNA complexes, we demonstrate the exist-

ence of numerous intermolecular contacts between the DNA major groove atoms C5(Cyt)/C5M(Thy) and protein oxygen atoms. A comprehensive analysis of the geometry of these close contacts exhibits features that are similar to other CH···O interactions found in proteins and small molecules, as well as to the NH···O hydrogen bonds observed in our data set. This similarity indicates that at least part of the CH···O contacts described here can be considered as hydrogen bonds.

As was suggested recently for protein structures (Derewenda *et al*., 1995), further refinement of protein–DNA complexes based on new potential functions, treating CH···O interactions as attractive rather than repulsive, would result in a better agreement of the CH···O bond geometries with the stereochemical features of classical hydrogen bonds. We propose that CH···O interactions, considered here for the first time in protein–DNA complexes, contribute to specific recognition of DNA target sites by proteins, and play a role in the modulation of gene expression. The functional advantage of these novel interactions may be related to their weakness. Indeed, the nucleo-protein recognition based entirely on strong hydrogen bonds, would make gene regulation hardly possible, due to the expected low dissociation rates of the complexes. The CH···O interactions not only facilitate attraction of the cognate sequences to each other, but at the same time make this attraction relatively weak, and therefore, more readily reversible. These interactions, being weaker and thus less selective compared with classical hydrogen bonds, are also likely to play an important role in the recognition of a range of DNA target sites by the same protein. Overall, including these CH···O interactions together with the conventional hydrogen bonds between amino acid side-chains and the DNA groove edges reveals more consistently the base–amino acid preferences, and has direct implications for molecular design experiments.

Methods

Geometric features of the short contacts ⩽3.5 Å

These were considered for selected interactions, including the CH···O and NH···O contacts (Table 2). Since in most cases the X-ray data do not provide the positions of hydrogen atoms, they were built according to strereochemical criteria. For all groups, the C–H bond length was defined as 1.08 Å and the N–H bond as 1.00 Å. The hydrogen for C5(Cyt) was positioned along the bisector of the angle C4-C5-C6, while for the amino groups N4(Cyt) and N6(Ade) the two hydrogen atoms were built in the plane of the base in sp^2 hybridization geometry (forming three angles of 120°). The three hydrogens of C5M(Thy) were placed as in ideal sp^3 (tetrahedral) hybridization; the first proton (with respect to which the other two hydrogens were positioned) was located as close as possible to the interacting protein oxygen, leaving a rotational freedom for the CH_3 group.

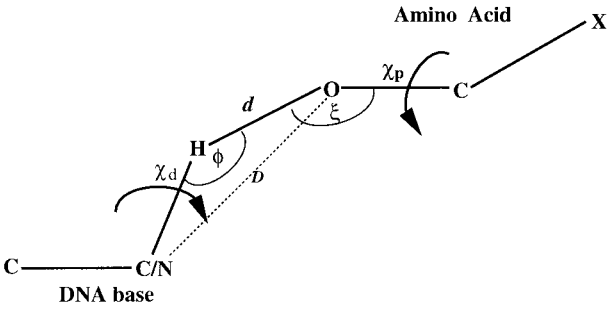

Figure 4. Geometric parameters characterizing the CH···O and NH···O interactions. The dihedral angles χ_d[C-C/N-H···O] and χ_p[H···O-C-X] are used to calculate the elevation angles θ_d and θ_p (see Methods). X stands for nitrogen in the case of interaction with the peptide carbonyl, and for carbon in the case of interaction with the side-chain oxygen (carbonyl in Asn and Gln, carboxyl in Asp and Glu, and hydroxyl in Ser, Thr and Tyr).

Each of the above contacts was characterized by six geometric parameters, illustrated in Figure 4. The first four parameters, the distance between the heavy atoms C/N···O (*D*), the length of the hydrogen bond H···O (*d*), the donor angle ϕ defined as C/N-H···O, and the acceptor angle ξ, defined as H-O-C(protein), are standard and frequently used for characterization of hydrogen bonds (Jeffrey & Saenger, 1991). In addition, two dihedral angles, χ_d and χ_p, defined, respectively, by atoms [C-C/N-H···O] and [H···O-C-X], were used to calculate the elevation angles θ_d and θ_p from the relationships: $\sin\theta_d = \sin\phi \sin\chi_d$; $\sin\theta_p = \sin\xi \sin\chi_p$. These notations and definitions are consistent with those introduced earlier by Taylor & Kennard (1982) and Derewenda *et al*. (1995) to analyze CH···O interactions in small molecules and proteins, respectively.

Normalized frequencies of C···O, N···O and C···C interactions

For a series of distances, ranging from 2.6 to 6.0 Å in increments of 0.2 Å , the number of contacts of each type was computed and normalized (Figure 3). For each distance, normalization was obtained by dividing the number of specific contacts at this distance by the total number of contacts between the corresponding DNA atom with all protein atoms. For example, the normalized number of C5(Cyt)···O interactions at the distance 3.0 Å is given by the number of pairs (C5, O) for those protein oxygen atoms that are at the distance of 2.8 to 3.2 Å from C5, divided by the number of interactions between C5(Cyt) and all protein atoms at the same distances. This kind of normalization is essential due to irregularities of the convex and concave shapes interacting at the protein–DNA interface.

Charges on the amino acid tips and DNA base edges

To calculate the net charges for the bases in the major groove, three different scales of atomic charges were considered (Renugopalakrishnan *et al*., 1971; Pearlman & Kim, 1990; Cornell *et al*., 1995) (Table 4). There is an almost complete agreement in the ranking of the bases by their charges according to all three scales used,

G < A ⩽ T < C. The only exception concerns the ranking of adenine *versus* thymine based on the charges of Cornell *et al*. (1995), according to which adenine is more positive than thymine by 0.1 (Table 4); however, the other two sets of partial charges predict that adenine is more negative than thymine by 0.1 to 0.3. Thus, we use the relationship A ⩽ T for ranking the bases in Table 3.

The amino acids are divided into several categories and arranged depending on their donor/acceptor propensity and the partial charges on their tips (Table 3). Evidently, the charged amino acids should be located at the opposite ends of this scale: Asp and Glu (denoted O^-) are the most negative and acceptor-prone, while Lys and Arg (denoted N^+) are the most positive and donor-prone. To rank the neutral amino acids (denoted OH/SH and O = /NH), which can serve both as donors and acceptors of protons, the net charges on their tips were calculated according to two sets of potential functions (Momany *et al*., 1975; Cornell *et al*., 1995). Notice that although the partial charges in the two sets differ quite substantially, the overall charges are consistent with each other (Table 5): in both cases side-chains in the (OH/SH) group are more negatively charged than in the (O = /NH) group. Thus, the four groups of amino acids are ranked in the following order according to the net charges on their tips: $(O^-) < (OH/SH) < (O = /NH) < (N^+)$.

Acknowledgments

We thank Ora Schueler and Dagmar Ringe for valuable discussions. This study was supported by the Israeli Science Foundation administered by the Israeli Academy of Sciences (granted to H.M.).

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 of high resolution. *Science,* **242**, 899–907.

Bahar, I. & Jernigan, R. L. (1997). Inter-residue potential in globular proteins and the dominance of highly specific hydrophilic interaction at close separation. *J. Mol. Biol.* **266**, 195–214.

Beamer, L. J. & Pabo, C. O. (1992). Refined 1. 8 Å crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.* **227**, 177–196.

Bella, J. & Berman, H. M. (1996). Crystallographic evidence for Cα-H···O = C hydrogen bonds in a collagen triple helix. *J. Mol. Biol.* **264**, 734–742.

Berger, I., Egli, M. & Rich, A. (1996). Inter-strand C-H···O hydrogen bonds stabilizing four-stranded intercalated molecules: stereoelectronic effects of O4′ in cytosine-rich DNA. *Proc. Natl Acad. Sci. USA,* **93**, 12116–12121.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **63**, 751–759.

Cheng, X., Balendiran, K., Schildkraut, I. & Anderson, J. E. (1994). Structure of *Pvu*II endonuclease with cognate DNA. *EMBO J.* **13**, 3927–3935.

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**, 346–355.

Choo, Y. & Klug, A. (1994). 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.

Cornell, W. D., Cieplak, P., Bayly, I. B., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). The occurrence of C-H···O hydrogen bonds in proteins. *J. Mol. Biol.* **252**, 248–262.

Desiraju, G. R. (1991). The C-H···O hydrogen bonds in crystals: what is it? *Acc. Chem. Res.* **24**, 290–296.

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.

Ellenberger, T. E., Brandl, C. J., Struhl, K. & Harrison, S. C. (1997). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterupted α helices: crystal structure of the protein-DNA complex. *Cell,* **71**, 1223–1237.

Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996). Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions. *Structure,* **4**, 1171–1180.

Fairall, L., Schwabe, J. W., 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.

Feng, J. A., Johnson, R. C. & Dickerson, R. E. (1994). Hin recombinase bound to DNA: the origin of specificity in major and minor groove interactions. *Science,* **263**, 348–355.

Ferré-D'Amaré, 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, 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.

Glover, J. N. & Harrison, S. C. (1995). Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature,* **373**, 257–261.

Hegde, R. S., Grossman, S. R., Laimins, 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.

Hirsch, J. A. & Aggarwal, A. K. (1995). Structure of the Even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. *EMBO J.* **14**, 6280–6291.

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

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Structures,* pp. 111–135, Springer, Berlin.

Kim, J. L. & Burley, S. K. (1994). 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nature Struct. Biol.* **1**, 638–653.

Kim, Y. C., Grable, J. C., Love, R., Greene, P. J. & Rosenberg, J. M. (1990). Refinement of *Eco*RI endonuclease crystal structure: a revised protein chain tracing. *Science,* **249**, 1307–1309.

Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993). Crystal structure of a yeast TBP/TATA-box complex. *Nature,* **365**, 512–520.

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.

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.

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**, 139–154.

Kostrewa, D. & Winkler, F. K. (1995). Mg2+ binding to the active site of *Eco*RV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 Å resolution. *Biochemistry,* **34**, 683–696.

Lawson, C. L. & Carey, J. (1993). Tandem binding in crystals of a *trp* repressor/operator half-site complex. *Nature,* **366**, 178–182.

Li, T., Stark, M. R., Johnson, A. D. & Wolberger, C. (1995). Crystal structure of the MATa1/MAT α2 homeodomain heterodimer bound to DNA. *Science,* **270**, 262–269.

Luisi, B. F., Xu, W., 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.

Lustig, B. & Jernigan, R. L. (1995). Consistencies of individual DNA base-amino acid interactions in structures and sequences. *Nucl. Acids Res.* **23**, 4707–4711.

Ma, P. C., 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.

Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA–complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.

Marmorstein, R., Carey, M., Ptashne, M. & Harrison, S. C. (1992). DNA recognition by GAL4: structure of a protein–DNA complex. *Nature,* **356**, 408–414.

Marmorstein, R. & Harrison, S. C. (1994). Crystal structure of a PPR1–DNA complex: DNA recognition by proteins containing a Zn2Cys6 binuclear cluster. *Genes Dev.* **8**, 2504–2512.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occuring amino acids. *J. Phys. Chem.* **79**, 2361–2381.

Mondragon, A. & Harrison, S. C. (1991). The phage 434 Cro/Or1 complex at 2. 5 Å resolution. *J. Mol. Biol.* **219**, 321–334.

Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L. & Harrison, S. C. (1995). Structure of the NF-κb p50 homodimer bound to DNA. *Nature,* **373**, 311–317.

Nikolov, D. B., Chen, H., Halay, E. D., Hoffman, A., Roeder, R. G. & Burley, S. K. (1996). Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl Acad. Sci. USA,* **93**, 4862–4867.

Ornstein, R. L. & Zheng, Y. (1997). *Ab Initio* quantum mechanics analysis of imidazole C-H···O water hydrogen bonding and a molecular mechanics forcefield correction. *J. Biomol. Struct. Dynam.* **14**, 657–665.

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.

Parkinson, G., Wilson, C., Gunasekera, A., Ebright, Y. W., Ebright, R. E. & Berman, H. M. (1996). Structure of the CAP–DNA complex at 2.5 Å resolution: a complete picture of the protein–DNA interface. *J. Mol. Biol.* **260**, 395–408.

Pauling, L. (1960). *The Nature of the Chemical Bond,* 3rd edit., Cornell University Press, Ithaca.

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.

Pearlman, D. A. & Kim, S. (1990). Atomic charges for DNA constituents derived from single-crystal X-ray diffraction data. *J. Mol. Biol.* **211**, 171–187.

Rao, B. J. & Radding, C. M. (1994). Formation of base triplets by non-Watson–Crick bonds mediates homologous recognition in RecA recombination filaments. *Proc. Natl Acad. Sci. USA,* **91**, 6161–6165.

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

Renugopalakrishnan, V., Lakshminarayanan, A. V. & Sasisekharan, V. (1971). Stereochemistry of nucleic acids and polynucleotides. III. Elecronic charge distribution. *Biopolymers,* **10**, 1159–1167.

Rodgers, D. W. & Harrison, S. C. (1993). The complex between phage 434 repressor DNA-binding domain and operator site OR3: structural differences between consensus and non-consensus half-sites. *Structure,* **1**, 227–240.

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**, 763–770.

Schwabe, J. W., 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. & Harrison, S. C. (1993). The phage 434 OR2/R1–69 complex at 2.5 Å resolution. *J. Mol. Biol.* **232**, 826–838.

Somers, W. S. & Phillips, S. E. (1992). Crystal structure of the met repressor–operator complex at 2.8 Å res-

olution reveals DNA recognition by beta-strands. *Nature,* **359**, 387–393.
Sponer, J., Leszczynski, J. & Hobza, P. (1996). Hydrogen bonding and stacking of DNA bases: a review of quantum-chemical *ab initio* studies. *J. Biomol. Struct. Dynam.* **14**, 117–135.
Sponer, J. & Kypr, J. (1994). Close mutual contacts of the amino groups in DNA. *Int. J. Biol. Macromol.* **16**, 3–6.
Sutor, D. J. (1962). The C-H···O hydrogen bond in crystals. *Nature,* **195**, 68–69.
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.
Taylor, R. & Kennard, O. (1982). Crystallographic evidence for the existence of C-H···O, C-H···N, and C-H···Cl hydrogen bonds. *J. Am. Chem. Soc.* **104**, 5063–5070.
Wahl, M. C., Rao, S. T. & Sundaralingam, M. (1996). The structure of r(UUCGCG) has a 5′-UU-overhang exhibiting Hoogsteen-like trans U. U base pairs. *Nature Struct. Biol.* **3**, 24–31.
Wahl, M. C. & Sundaralingam, M. (1997). C-H···O hydrogen bonding in biology. *TIBS,* **22**, 97–102.
Wilson, D. S., Guenther, B., Desplan, C. & Kuriyan, J. (1995). High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell,* **82**, 709–719.
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.
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**, 639–650.
Zhurkin, V. B., Raghunathan, G., Ulyanov, N. B., Camerini, Otero R. D. & Jernigan, R. L. (1994). A parallel DNA triplex as a model for the intermediate in homologous recombination. *J. Mol. Biol.* **239**, 181–200.

Edited by B. Honig

(Received 5 August 1997; received in revised form 20 January 1998; accepted 21 January 1998)