Structural Features of Protein−Nucleic Acid Recognition Sites†

Katalin Nadassy,‡,§ Shoshana J. Wodak,‡,|| and Joël Janin*,‡,⊥

European Bioinformatics Institute, EMBL, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England, Department of Biological and Molecular Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UCMB, Université Libre de Bruxelles, CP 160/16, Avenue P. Héger, 1050-Bruxelles, Belgium, and Laboratoire d'Enzymologie et de Biochimie Structurales, CNRS UPR9063, 91198-Gif-sur-Yvette, France

Received October 2, 1998; Revised Manuscript Received November 30, 1998

ABSTRACT: We analyzed the atomic models of 75 X-ray structures of protein−nucleic acid complexes with the aim of uncovering common properties. The interface area measured the extent of contact between the protein and nucleic acid. It was found to vary between 1120 and 5800 Å^2. Despite this wide variation, the interfaces in complexes of transcription factors with double-stranded DNA could be broken up into recognition modules where 12 ± 3 nucleotides on the DNA side contact 24 ± 6 amino acids on the protein side, with interface areas in the range 1600 ± 400 Å^2. For enzymes acting on DNA, the recognition module is on average 600 Å^2 larger, due to the requirement of making an active site. As judged by its chemical and amino acid composition, the average protein surface in contact with the DNA is more polar than the solvent accessible surface or the typical protein−protein interface. The protein side is rich in positively charged groups from lysine and arginine side chains; on the DNA side the negative charges from phosphate groups dominate. Hydrogen bonding patterns were also analyzed, and we found one intermolecular hydrogen bond per 125 Å^2 of interface area in high-resolution structures. An equivalent number of polar interactions involved water molecules, which are generally abundant at protein−DNA interfaces. Calculations of Voronoi atomic volumes, performed in the presence and absence of water molecules, showed that protein atoms buried at the interface with DNA are on average as closely packed as in the protein interior. Water molecules contribute to the close packing, thereby mediating shape complementarity. Finally, conformational changes accompanying association were analyzed in 24 of the complexes for which the structure of the free protein was also available. On the DNA side the extent of deformation showed some correlation with the size of the interface area. On the protein side the type and size of the structural changes spanned a wide spectrum. Disorder-to-order transitions, domain movements, quaternary and tertiary changes were observed, and the largest changes occurred in complexes with large interfaces.

Protein−nucleic acid recognition plays an essential role in all mechanisms of gene expression and control. Its biochemical and structural basis has been a field of intense study in the last 10 years (*1*−*2*), and it is now firmly established thanks to a large number of recent X-ray and NMR[1] structures of protein−DNA and protein−RNA complexes. Individual complexes have been analyzed and described in atomic details by authors of these structures, and a few related systems have been compared to each other. Several studies were devoted to analyzing the geometrical principles and interaction trends in DNA complexes with enzymes and transcription factors (*3*−*5*). Yet, general surveys have been rare (*6*−*9*) and their conclusions need updating in face of a large body of new data.

Here, we examine a sample of 75 X-ray structures of protein−nucleic acid complexes, mostly with double-stranded DNA, and attempt to identify common properties and rules that apply across complexes irrespective of the specific systems involved. To that end we evaluate the extent of the contact between the protein and nucleic acid by computing the interface area in the complex. We analyze the nature of the atoms, amino acids, and nucleotide residues that constitute the interface, review the polar interactions that they make, and examine interactions with solvent. We measure the volume of protein atoms at the interfaces using Voronoi polyhedra, and compare it to the volume of atoms buried inside proteins, to assess how well the protein−nucleic acid interfaces are packed. We give a general description of the conformation changes taking place in the protein and in the nucleic acid components of the complexes. Last, we compare protein−nucleic acid to protein−protein recognition, and observe that several features of complexes involving DNA and proteins as diverse as endonucleases and transcription factors have an equivalent protein−protein recognition where both components are polypeptides. For instance, a recognition module with an interface area of 1600 ± 400 Å^2 can be identified in both types of recognition processes. These features may therefore characterize the interaction of biological macromolecules in general.

† K.N. was supported by a studentship of University of Stirling, J.J. by Université Paris-Sud during a sabbatical leave. S.W. thanks the EU Biotech program for supporting part of this work.

* Corresponding author: Joël Janin, Laboratoire d'Enzymologie et de Biochimie Structurales, CNRS UPR9063, 91198-Gif-sur-Yvette, France. E-mail: janin@lebs.cnrs-gif.fr. Fax: (33) 1 69 82 31 29.

‡ EMBL.

§ University of Stirling.

|| Université Libre de Bruxelles.

⊥ CNRS.

[1] Abbreviations: NMR, nuclear magnetic resonance; PDB, Protein Data Bank; TBP, TATA box binding protein; RMSD, root-mean-square distance.

10.1021/bi982362d CCC: $18.00 © 1999 American Chemical Society
Published on Web 01/28/1999

Protein−Nucleic Acid Complexes of Known Structure. At the end of 1997, over 200 entries in the Brookhaven Protein Data Bank (PDB; ref *10*) described protein−nucleic acid complexes. These entries were defined as containing proteins and oligonucleotides at least four nucleotide units long. Among them, we selected 75 X-ray structures having at least 3.2 Å resolution (Table 1). All but three have a resolution of 3.0 Å or better, and twenty-eight, 2.4 Å or better. Sixty-five contain double-stranded DNA, four, single-stranded DNA, and six, single-stranded RNA. Of the 65 complexes with double-stranded DNA, 19 are with proteins involved in DNA replication, hydrolysis, recombination, modification, or repair, among which 16 are enzymes. The remaining ones are all transcription factors: 9 from bacteria or bacteriophages and 37 from eukaryotes.

Because most entries include more than one polypeptide or nucleic acid chain, biologically significant assemblies had to be selected from the PDB entries. The selection was made as described in Table 1. In addition to the PDB code of each entry, the table lists codes for polypeptide or nucleic acid chains followed by the number of amino acids or nucleotide residues in the chain. The process of selection was not without ambiguities. Take the *Escherichia coli* lactose repressor−operator interaction as an example. It is represented in the PDB by the NMR structure of the 6 kDa N-terminal headpiece of the repressor bound to a 11-bp half-operator (1lcc, ref *11*), and by a low-resolution X-ray structure of the whole molecule, a 160 kDa tetramer bound to two copies of a 21-bp symmetric operator (1lbg, ref *12*). Whereas the headpiece retains affinity and some specificity for the operator DNA, several contacts with DNA are lost in the protein fragment relative to the whole polypeptide chain, and all are repeated four times in the tetramer. The high-resolution X-ray structure of the dimeric purine repressor bound to the symmetric Pur F operator (1wet) was retained as a representative of this system, for it is homologous to the lactose repressor and the dimer binds its operator in the same way as each half of the tetramer.

Like the lactose repressor headpiece, most eukaryotic transcription regulators appear in the PDB as fragments of much larger polypeptide chains. The oligomeric state of the fragment may differ from that of the whole protein and depend on the presence or absence of DNA. It sometimes changes with the sequence and length of the oligonucleotide in the cocrystal. We retained the dimer as being the biological unit in the case of several homeodomains (1apl, 1fjl) and hormone receptor DNA-binding domains (1glu, 1hcq). They are monomeric in solution, but the protein−protein contacts seen in the X-ray structures are considered to be biologically significant. Somewhat arbitrarily, we also treated as dimers several bacterial repressors (1cma, 1par, 1tro) that can be viewed either as dimers or tetramers. Finally, yeast aspartyl-tRNA synthetase (1asy) was treated as a monomer and seryl-tRNA synthetase (1ser) as a dimer even though both are homodimers in solution, because in the first enzyme, each of the two subunits interacts independently with a tRNA molecule, whereas the X-ray structure of the second enzyme shows a single tRNA interacting with both subunits. In these two cases, we retained a protein−RNA interface rather than the structural unit.

Due to the uncertain nature of the biologically relevant assembly, we also defined *binding units* for the protein component of each of the complexes. In most cases, the binding unit is a polypeptide chain that occurs either as a monomer or as part of a dimer of identical or closely related chains. Exceptions are assemblies of polypeptide chains with unrelated folds (the TATA box binding protein associated with other components of the transcription machinery) which were counted as a single unit, and polypeptide chains containing tandem repeats (the Cys_2His_2 zinc finger proteins, some of the homeodomains), where each repeat was counted as a binding unit if it actually bound DNA. The number of binding units in each complex is listed in Table 2.

Size of Protein−Nucleic Acid Interfaces. Interface Area and Interface Residues. The extent of the contact between two macromolecules is measured as the interface area *B* in Table 2. This can be derived from atomic coordinates of the complex by computing its solvent accessible surface area and subtracting it from the sum of the accessible surface areas of the isolated components. Accessible surface areas were evaluated with program SURVOL (*13*), which implements the Lee and Richards (*14*) algorithm. Group radii were from ref *15*, and the radius of the water probe was 1.5 Å. *B* measures the area of protein and nucleic acid surface that is buried in contacts between the two macromolecules. Because the component structures are taken from the complex, *B* may differ from the area of the protein/nucleic acid accessible surface that is lost upon association in cases where conformation changes accompany association. Interface residues are defined as amino acids or nucleotides that lose accessible surface in the complex. Their number N_{aa} and N_{nuc} in each complex is given in Table 2 and their average number in classes is given in Table 3. Figure 1 shows that the interface area *B* and the number of interface residues are two highly correlated measures of the size of the interface (correlation coefficient 0.9).

B is defined here as the area of the surface buried on both the protein and the nucleic acid. The two components contribute equally to *B* on average, and also in most individual complexes. The fraction of *B* contributed by the protein is 50% ± 5% for all but two complexes: polymerase-β (1bpy) and the barnase-tetranucleotide complex (1brn). There, the fraction is near 40% and the protein loses significantly less accessible surface than the nucleic acid. While the contributions to *B* are generally the same, there are more amino acid than nucleotide residues at the interfaces. Because of its smaller size, the average interface amino acid loses less accessible surface than the average nucleotide. The ratio N_{aa}/N_{nuc} of the number of interface amino acid/nucleotide residues is ≈2.0, and the ratio B_{aa}/B_{nuc} of their contributions to *B* is 0.5 in complexes with double-stranded DNA. The first ratio increases to 2.5 in complexes with RNA and to 4.5 with single-stranded DNA. On average, an interface nucleotide loses approximately 40% of its solvent accessible surface area when a protein binds to double-stranded DNA or RNA and almost twice as much when it binds to single-stranded DNA.

The Range and Distribution of Interface Areas. The size of the interface varies widely from one complex to another: the range is 1120−5800 Å^2 for *B*, 18−90 for N_{aa}, and 10−49 for N_{nuc} not counting a tetranucleotide bound to the small bacterial ribonuclease barnase (1brn). Values of *B* listed in Table 2 are plotted as a histogram in Figure 2a. Three-quarters of the values are in the range 3000 ± 1200 Å^2, yet

Table 1. Protein-Nucleic Acid Complexes

code	protein	res (Å)	chain code and number of amino acids	chain code and number of nucleotides	reference
A. Double-stranded DNA (*65*)					
DNA replication, hydrolysis, recombination, modification, and repair (*19*)					
DNA polymerases					
1bpy	DNA polymerase β, human	2.2	A335	T16, P10, D5	48
1kln	Klenow fragment, *E. coli* Pol I	3.2	A605	B23	49
1tau	Taq, *T. aquaticus*	3.0	A832	T8, P8	50
1t7p	phage T7	2.2	A698, B108	P11, T13	51
endonucleases					
1bhm	*Bam*HI	2.2	A213, B213	C12, D12	33
1dnk	DNase I, bovine	2.3	A260	B7, C8	52
1eri	*Eco*RI	2.7	A276[a]	B13[a]	53
1fok	*Fok*1	2.8	A568	B20, C20	18
1pvi	*Pvu*II	2.8	A157, B157	C13, D13	54
1rvc	*Eco*RV	2.1	A244, B244	C6, D5, E6, F5	55
1vas	phage T4 endonuclease V	2.8	A137	B13, C13	56
other enzymes					
1dct	*Hae*III cytosine methylase	2.8	A324	F18, M18	43
1gdt	$\gamma\delta$-resolvase	3.0	A183, B183	C22, D13, E21, F13	20
1hcr	*Hin* recombinase	2.3	A52	B13, C14	57
1mht	*Hha*I methyl transferase[b]	2.8	A327	B12, C12	58
1tc3	*C. elegans* transposase	2.4	C51	A21, B20	59
miscellaneous					
1ecr	replication terminator, *E. coli*	2.7	A309	B16, C16	60
1ign	RAP1 telomere binding, yeast	2.3	A246	C19, D19	61
1ihf	integration host factor *E. coli*	2.5	A99, B94	C35, D15, F20	62
prokaryotic transcription factors (*9*)					
1cma	Met J repressor, *E. coli*[c]	2.8	A104, B104	C10, D9	63
1lmb	repressor, phage lambda	1.8	*3*:92, *4*:92	*1*:20, *2*:20	34
1par	Arc R, phage P22[c]	2.6	A53, B53	E22, F22	64
1per	repressor, phage 434	2.5	L69, R69	A20, B20	65
1ruo	CAP, *E. coli*	2.7	A209, B209	C14, D17, E14, F17	66
1tro	Trp repressor, *E. coli*[c]	1.9	A108, C108	I19, J19	29
1trr	Trp repressor, half-site DNA	2.4	A107, B107	C16, I16	67
1wet	purine repressor, *E. coli*	2.6	A340[a]	B17[a]	68
3cro	Cro repressor, phage 434	2.5	L71, R71	A20, B20	69
eukaryotic transcription factors (*37*)					
homeodomains					
1apl	Mat-α2, yeast	2.7	C83, D83	A21, B21	70
1fjl	paired dimer, *Drosophila*	2.0	A81, B81	D14, E14	71
1hdd	engrailed, *Drosophila*	2.8	C61	A21, B21	72
1oct	Oct-1 POU	3.0	C156	A15, B15	73
1pdn	paired, *Drosophila*	2.5	C128	A15, B15	74
1yrn	Mat-A2/Mat-α2, yeast	2.5	A61, B83	C42	75
leucine zippers, bHLH, ETS					
1an2	max, human	2.9	A86, C86	B22, C22	76
1an4	USF	2.9	A65, B65	C21, D21	77
1fos	c-Fos/c-Jun dimer	3.1	E60, F57	A20, B20	78
1hlo	max (intact dimer), human	2.8	A150, B150	C13, D13	79
1mdy	Myo D	2.8	A68, B62	E14, F14	80
1pue	Pu1-ETS domain	2.1	E89	A16, B16	81
1srs	serum response factor	3.2	A92, B92	W19, C19	82
1ysa	GCN4, yeast	2.9	C57, D57	A20, B20	83
2dgc	GCN4,yeast (ATF site)	2.2	A63[a]	B19[a]	84
zinc fingers					
1aay	Zif 268	1.6	A90	B11, C11	85
1mey	designed	2.2	C87	A13, B13	86
1ubd	YY1	2.5	C124	A20, B20	87
2drp	tramtrack	2.8	A66	B19, C19	88
2gli	GLI1	2.6	A160	C21, D21	89
other zinc modules					
1d66	Gal4	2.7	A66, B66	D19, E19	90
1glu	glucocorticoid receptor[c]	2.9	A81, B81	C19, D19	91
1hcq	estrogen receptor[c]	2.4	A84, B84	C18, D18	92
1lat	glucocorticoid receptor (noncognate DNA site)	1.9	A82	C19, D19	93
1pyi	pyrimidine pathway regulator	3.2	A96, B96	D14, E14	94
1zme	proline utilization PUT3	2.5	C70, D70	A17, B17	95
2nll	retinoid receptor	1.9	A66, B103	C18, D18	96
TATA box binding protein					
1ais	TBP-TFIIB, *Pyrococcus*	2.1	A182, B200	C17, E17	97
1cdw	TBP,human	1.9	A179	B16, C16	98
1vol	TBP-TFIIB, *Arabidopsis*	2.7	A204, B200	C16, D16	99
1ytf	TBP-TFIIA, yeast	2.5	A180, B53, C79, D121	E16, F16	100

Table 1. (Continued)

code	protein	res (Å)	chain code and number of amino acids	chain code and number of nucleotides	reference
others					
1a3q	NFκ-B p52, human	2.1	A285, B285	C11, D11	101
1nfk	NFκ-B p50, mouse	2.3	A325, B325	C11, D11	102
1svc	NFκ-B p50, human	2.6	P365[a]	D19[a]	103
1tsr	p53 core	2.2	B219	E21, F21	16
1xbr	T-domain, *Xenopus*	2.5	A184, B184	C24, D24	104
2bop	E2 domain, papillomavirus-1	1.7	A85[a]	B17[a]	105
B. single-stranded DNA complexes (*4*)					
1brn	barnase	1.8	L110	A4	106
1hut	thrombin	2.9	L36, H259	D15	107
1jmc	replication protein RPA70	2.4	A246	B8	108
1uaa	rep helicase	3.0	A673, B673	C16	109
C. RNA complexes (*6*)					
1asy	Aspartyl-tRNA synthetase[c]	3.0	A490	B75	110
1gtr	Glutaminyl-tRNA synthetase	2.5	A553	B74	111
1ser	Seryl-tRNA synthetase	2.9	A421,B421	T94	112
1ttt	elongation factor EF-Tu	2.7	A405	D77	113
1urn	spliceosome U1A protein	1.9	A97	P21	114
1zdi	phage MS2	2.7	A129	R19	115

[a] The dimer was generated by applying a crystal symmetry operation. [b] The protein is covalently bonded to the DNA. [c] The oligomeric state of the protein is discussed in the text.

the histogram is clearly multimodal. At the lower end, there is an isolated group of six interfaces with *B* less than 1600 Å^2, three of which are less than 1200 Å^2. The three smallest interfaces are for a fragment of RNA bound to the capsid of phage MS2 (1zdi), the tetranucleotide bound to barnase (1brn), and a short piece of single-stranded aptamer DNA bound to the serine protease thrombin (1hut). The latter two do not represent biologically relevant protein−nucleic acid interactions.

The other three complexes in this group contain double-stranded DNA and have *B* values near 1400 Å^2. They contain the DNA-binding domain of the p53 oncogene protein (1tsr), DNase I (1dnk), and the DNA-binding domain of the glucocorticoid receptor bound to a noncognate DNA site (1lat). Their interfaces involve 18−33 amino acid and 10−14 nucleotide residues. The extent of the contact DNase I makes with its DNA substrate is probably underestimated, for the oligonucleotide crystallized with it is an octamer, and the structure suggests that a longer DNA would make additional contacts. When the glucocorticoid receptor binds cognate instead of noncognate DNA (1glu), it dimerizes and the interface area doubles. The crystal structure of the p53 DNA-binding domain−DNA complex contains a 21-bp oligonucleotide and three independent copies of the domain. The contact analyzed here is between one of the domains and the central part of the DNA. A second domain is in contact with its extremities, and the third domain does not contact DNA at all (*16*). This suggests that the affinity of the domain for DNA is marginal, in line with the small interface it can form. In the complete p53 protein, which is a tetramer, the interface is expected to be at least four times larger.

The other 69 complexes have $B > 1800$ Å^2. The largest values exceed 5000 Å^2 in $\gamma\delta$-resolvase (1gdt), where 76 amino acid residues contact 49 nucleotides, in the *E. coli* replication terminator (1ecr), the integration host factor (1ihf), and between glutaminyl aminoacyl-tRNA synthetase and its cognate tRNA (1gtr).

Interfaces in Binding Units: The Recognition Module. Most of the very large interfaces occur in complexes containing homooligomeric or tandem-repeat proteins, which can be considered as made of several binding units as described above. This is analyzed in the histogram of Figure 2b. It differs from that of Figure 2a in that it only includes double-stranded DNA complexes and displays the interface area per binding unit (*B*/unit) instead of *B*. The smaller interfaces now cover less than 1000 Å^2. They concern the zinc fingers, the *E. coli* Met J repressor (1cma), and the yeast PPR1 pyrimidine pathway regulator (1pyi). The case of the zinc fingers is remarkable, for this classical DNA-binding motif always occurs in more than one copy. In Figure 3, we plot the interface area as a function of the number of motifs, having cut the three-finger Zif 268 fragment (1aay) into single units or into pairs. The plot is linear with a slope of 900 Å^2 per unit, which fits with the value of B, observed in two other structures (1mey, 2gli). The slope indicates that a zinc finger−DNA interface contains about 13 amino acid residues losing 450 Å^2 of accessible surface area and about 7 nucleotides that lose an equivalent amount. This is probably the maximum that a polypeptide of 30 residues, the typical size of a zinc finger, can do.

Not counting the zinc fingers, half of the complexes (31 out of 60) have *B*/unit in the range 1600 ± 400 Å^2. This range covers the large majority of the transcription factors, but no enzyme other than DNase I. In comparison with the histogram of *B* in Figure 2a, the histogram of *B*/unit for the transcription factors in Figure 2b is fairly narrow. It expresses the fact that many DNA-binding proteins contain substructures which form an interface of a defined size and which we shall call *recognition modules*. Recognition modules comprise protein and DNA. They have a different structure in different complexes, but they have common features. On the protein side, a recognition module contributes 24 ± 6 amino acid residues to the interface losing 800 ± 200 Å^2 of accessible surface. Cys_2His_2 zinc fingers are too small to achieve that; other classical DNA-binding motifs can. On the DNA side, a module contains 12 ± 3 nucleotides that also lose 800 ± 200 Å^2. Some nucleotides occur as base pairs, but others may be distributed along each of the two strands of the double helix.

Table 2. Protein−Nucleic Acid Interfaces

code	protein	binding units	interface residues N_{aa}	interface residues N_{nuc}	interface area B (Å^2)	H-bonds N_{hb}	waters N_{wat}[a]
A. Double-stranded DNA							
DNA replication, hydrolysis, recombination, modification, and repair							
DNA polymerases							
1bpy	DNA polymerase β	1	59	17	2990	24	25
1kln	Klenow fragment	1	55	16	2740	18	
1tau	Taq, *T. aquaticus*	1	48	14	2530	10	
1t7p	phage T7	1	75	22	4000	23	21
endonucleases							
1bhm	*Bam*HI	2	79	21	4270	37	38
1dnk	DNase I	1	33	10	1530	13	6
1eri	*Eco*RI	2	76	24	4120	34	
1fok	*Fok*1	1	68	22	3910	28	
1pvi	*Pvu*II	2	82	24	4550	27	
1rvc	*Eco*RV	2	90	22	4870	37	37
1vas	phage T4 endonuclease V	1	45	17	2780	16	
other enzymes							
1dct	*Hae*III cytosine methylase	1	52	20	3130	31	
1gdt	$\gamma\delta$-resolvase	2	76	49	5810	30	
1hcr	*Hin* recombinase	1	26	22	2880	15	2
1mht	*Hha*I methyl transferase[b]	1	57	17	3300	36	
1tc3	*C. elegans* transposase	1	34	19	2010	26	3
miscellaneous							
1ecr	replication terminator	1	78	30	5300	20	
1ign	RAP1 telomere binding	2	60	33	4410	47	26
1ihf	integration host factor	2	77	47	5120	38	
prokaryotic transcription factors							
1cma	Met J repressor	2	30	16	1890	20	
1lmb	lambda repressor	2	51	27	3080	28	20
1par	Arc R	2	33	18	2090	21	
1per	434 repressor	2	43	26	2940	20	
1ruo	CAP	2	45	34	3070	19	
1tro	Trp repressor	2	47	30	3250	26	28
1trr	Trp repressor, half-site	2	50	23	3060	24	22
1wet	purine repressor	2	58	28	3820	20	
3cro	434 Cro repressor	2	48	28	3040	20	
Eukaryotic transcription factors							
homeodomains							
1apl	Mat-α2	2	41	35	3740	19	
1fjl	paired dimer	2	45	26	3640	35	43
1hdd	engrailed	1	22	16	1890	12	
1oct	Oct-1 POU	2	50	23	3360	23	
1pdn	paired	2	40	22	2690	20	
1yrn	Mat-A2/Mat-α2	2	42	32	3550	22	
leucine zippers, bHLH, and others							
1an2	Max	2	32	24	2740	12	
1an4	USF	2	43	29	3020	13	
1fos	c-Fos/c-Jun dimer	2	30	21	2330	8	
1hlo	Max intact dimer	2	35	20	2700	14	
1mdy	Myo D	2	36	22	2820	19	
1pue	Pu1-ETS domain	1	29	19	2160	14	13
1srs	serum response factor	2	48	33	4200	27	
1ysa	GCN4	2	30	22	2400	18	
2dgc	GCN4, ATF site	2	32	20	2620	26	10
zinc fingers							
1aay	Zif 268	3	41	22	2870	20	35
1mey	designed	3	44	24	2670	24	26
1ubd	YY1	4	50	26	3030	14	
2drp	Tramtrack	2	28	16	1800	14	
2gli	GLI1[c]	4	57	30	3420	22	
other zinc modules							
1d66	Gal4	2	37	24	2630	24	
1glu	glucocorticoid receptor	2	51	24	2880	12	
1hcq	estrogen receptor	2	47	22	2630	31	21
1lat	glucocorticoid (noncognate)	1	23	13	1410	14	12
1pyi	pyrimidine pathway regulator	2	24	20	1980	8	
1zme	proline utilization PUT3	2	38	27	2860	23	
2nll	retinoid receptor	2	55	26	3380	29	29
TATA box binding protein							
1ais	TBP-TFIIB, *Pyrococcus*	1	68	25	3790	19	17
1cdw	TBP, human	1	46	19	3020	13	18
1vol	TBP-TFIIB, *Arabidopsis*	1	66	25	4090	19	
1ytf	TBP-TFIIA, yeast	1	52	20	3430	15	

Table 2. (Continued)

code	protein	binding units	interface residues N_{aa}	interface residues N_{nuc}	interface area B (Å^2)	H-bonds N_{hb}	waters N_{wat}[a]
others							
1a3q	NFκ-B p52, human	2	53	22	3300	22	19
1nfk	NFκ-B p50, mouse	2	59	22	3750	22	20
1svc	NFκ-B p50, human	2	66	28	4360	26	
1tsr	p53 core	1	18	14	1250	11	7
1xbr	T-domain	2	68	32	4480	23	
2bop	E2 domain	2	50	30	3300	24	28
B. single-stranded DNA complexes							
1brn	barnase	1	22	4	1110	15	12
1hut	thrombin	1	14	7	1030	2	
1jmc	RPA70	1	43	8	2270	12	4
1uaa	Rep helicase	1	47	10	2490	9	
C. RNA complexes							
1asy	aspartyl-tRNA synthetase	1	82	31	4670	27	
1gtr	glutaminyl-tRNA synthetase	1	86	35	5650	58	
1ser	Seryl-tRNA synthetase	1	42	24	2470	13	
1ttt	elongation factor EF-Tu	1	57	19	3110	20	
1urn	spliceosome U1A	1	28	12	1810	21	15
1zdi	Phage MS2	1	30	13	920	8	

[a] Solvent molecules within 3.5 Å of atoms of both components in 28 complexes with resolution 2.4 Å or better. [b] The protein is covalently bonded to the DNA. [c] The fragment contains five zinc fingers, but only four contact DNA

Table 3: Interface Amino Acid and Nucleotide Residues[a]

complex	interface area B (Å^2)	amino acids N_{aa}	amino acids B_{aa} (Å^2)	nucleotides[b] N_{nuc}	nucleotides[b] B_{nuc} (Å^2)
double-stranded DNA (*65*)					
mean	3180	49	34	24	68
sd	940	17	7	7	16
Single-stranded DNA (*4*)					
mean	1720	32	27	7	130
sd	760	16	7	3	41
RNA complexes (*6*)					
mean	3100	52	29	21	74
sd	1800	28	2	11	12
all complexes (*75*)					
mean	3100	48	33	23	72
sd	1050	18	7	8	23

[a] B, interface area and N_{aa} and N_{nuc}, number of interface amino acid and nucleotide residues, are from Table 2. B_{aa} and B_{nuc} are the average contribution to B of an amino acid or nucleotide; sd, standard deviation. [b] The average accessible surface area of a nucleotide is 172 Å^2 in free double-stranded DNA.

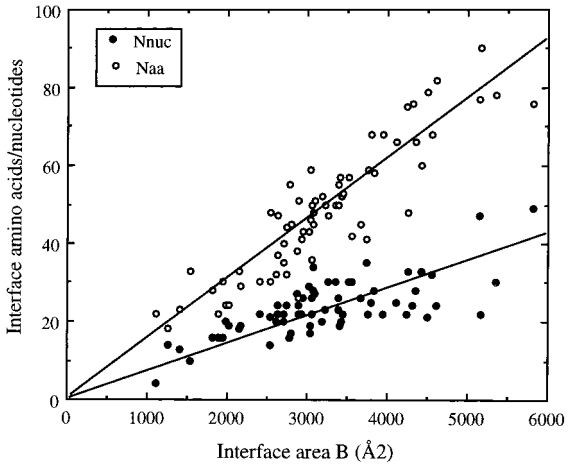

FIGURE 1: Correlation between the number of interface residues and the interface area in protein−nucleic acid complexes. All nucleotides (full circles) and amino acids (empty circles) that lose accessibility are counted as interface residues.

Recognition modules occur isolated in a few cases, or more often as a pair, in DNA complexes with prokaryotic transcription factors containing the classical helix-turn-helix motif, with homeodomains, with various types of leucine zippers, with Gal4 and hormone receptor-type zinc modules, and with NFκ-B and other eukaryotic transcription factors of the Rel family. Their presence can also be recognized in complexes with the TATA box binding protein (TBP). Though it is a single-chain protein and the complex buries over 3000 Å^2, TBP has an internal duplication not seen in the sequence (*17*). Its interface would comply with the 1600 $\pm$ 400 Å^2 rule if we counted TBP as two units. There is, however, no evidence that a half-TBP can be isolated or that it would bind DNA.

For enzymes, the range of B/unit that defines a recognition module must be modified. Complexes with five of the restriction endonucleases have values of B/unit in the range 2200 $\pm$ 250 Å^2. This suggests that the requirements to make an active site add about 600 Å^2 to the value of B/unit found in transcription factors. However, the complex with Fok1, which is monomeric, buries much more area, nearly 4000 Å^2. Fok1 is made of two distinct domains, a DNA recognition and a catalytic domain (*18*), and the contribution to B of contacts made by the DNA recognition domain is in the 1600 $\pm$ 400 Å^2 range. The four complexes with DNA polymerases bury 2500−4000 Å^2, yet they are single-chain proteins with no internal repetition. Moreover, as almost all nucleotides present in the crystal are in contact with the polymerase, the value of B in these complexes correlates with the size of the DNA fragment. We should expect it to reach even greater values with longer primer or product polynucleotides. DNA polymerases are multidomain proteins shaped like a hand. They contain at least the three domains called "fingers", "palm" and "thumb" that contact DNA, the palm domain carrying the polymerase active site (*19*). As in TBP, each domain could be considered as part of a different recognition module, but it is unlikely to be stable and functional by itself.

The distinction between a DNA recognition and a catalytic domain in an enzyme is best illustrated by $\gamma\delta$-resolvase (1gdt). DNA recognition in this recombination enzyme is

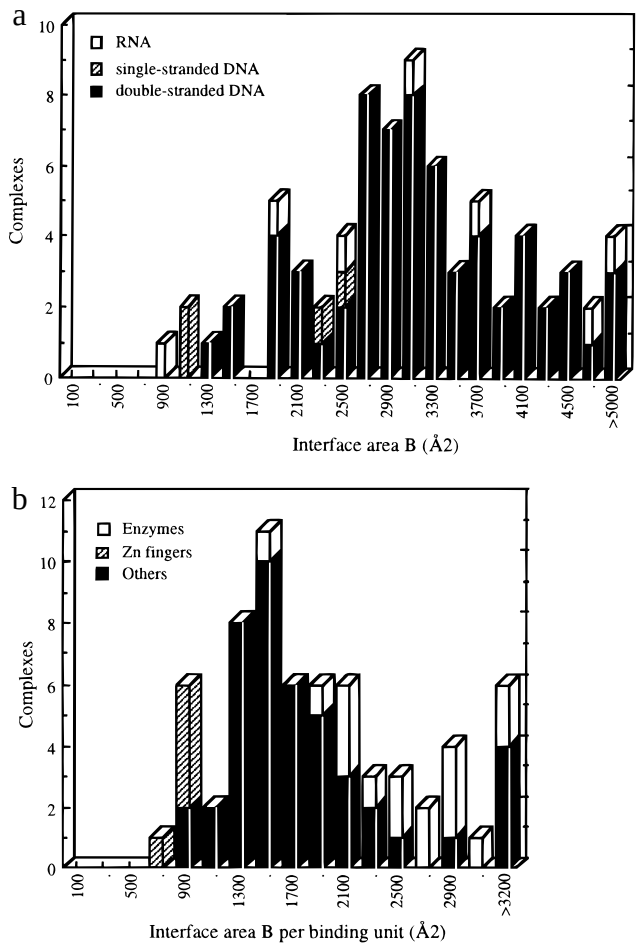

FIGURE 2: Histogram of interface areas. Distribution of observed values of (a) *B* in the 75 protein−nucleic acid complexes; and (b) *B*/unit in the binding units of the 65 protein−double-stranded DNA complexes. The number of binding units of each complex is quoted in Table 2.

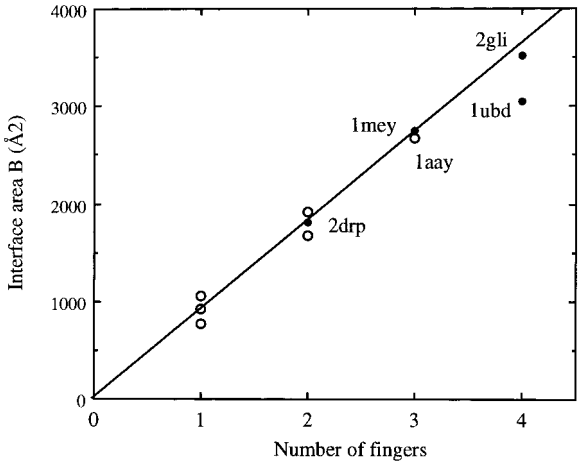

FIGURE 3: Interface area of zinc finger−DNA complexes: empty circles, values of *B*/unit for isolated zinc fingers or pairs of zinc fingers taken from the Zif 268 protein (1aay); and full circles, values for other zinc finger proteins.

achieved by a C-terminal segment containing a long α-helix followed by a globular domain of about 40 residues (*20*). The globular domain contains a helix-turn-helix motif and can be separated from the catalytic domain by limited proteolysis. γδ-Resolvase is a dimer, and each of two globular domains of the dimer accounts for 1900 Å^2 of the 5800 Å^2 buried in contacts with DNA. The globular domain therefore complies with the 1600 ± 400 Å^2 rule defining the recognition module. Other very large interfaces in Table 1 involve multidomain proteins that contact DNA or RNA through several domains. The *E. coli* replication terminator or the aminoacyl−tRNA synthetases are examples, but they are not as obviously modular as the transcription factors and γδ-resolvase.

Table 4: Chemical Character of Double-Stranded DNA−Protein Interfaces

type of surface[a]	buried at interface (%)	accessible in complex (%)
DNA		
phosphate[b]	43 (8)	35 (4)
sugar	29 (12)	38 (4)
base	27 (7)	28 (3)
non polar	41 (7)	47 (3)
neutral polar	16 (4)	19 (3)
charged (negative)	43 (8)	34 (4)
protein		
main chain	13 (7)	20 (4)
side chain	87 (7)	80 (4)
nonpolar	52 (8)	56 (3)
neutral polar	24 (7)	23 (4)
charged (positive)	23 (9)	12 (3)
(negative)	2 (2)	9 (3)

[a] Average fraction of the interface area contributed by either protein or DNA atoms. The standard deviation is in parentheses. All carbon-containing groups are counted as nonpolar, heteroatoms as polar. [b] O1P, O2P, and P atoms only; the two phosphodiester oxygens are attributed to the sugar moiety.

Chemical Composition of the Interfaces. Atomic Composition. The average chemical composition of interfaces in the 65 double-stranded DNA−protein interfaces is given in Table 4 and compared to that of the surface that remains accessible to the solvent. The compositions are given as fractions of the interface or accessible surface areas. Similar data could be derived for single-stranded DNA and RNA complexes, but their small number makes averaging meaningless.

On the protein side, the main chain contributes less to interfaces than to the average protein surface. Though the main chain contribution is not negligible, 87% of the protein contribution is from side chains. The hydrophilic/hydrophobic character of the surfaces in contact can be estimated by breaking the interface area into a nonpolar component (aliphatic and aromatic carbon atoms), a neutral polar component, and a component that bears full electric charges. The average surface of oligomeric proteins is 57% nonpolar, 22% neutral polar, and 21% charged (*21*). The surface composition of the protein component of the 65 complexes is close to that average. Compared to these solvent accessible surfaces, the surface in contact with DNA has a smaller nonpolar component (52%), a similar neutral polar component (24%), and a larger charged component (25%).

Moreover, there is an obvious difference in the charged surface component: the sign of the charge. On the average protein surface, positive and negative charges approximately equilibrate. Interfaces with DNA are highly enriched in positive charges from lysine and arginine side chains and almost entirely devoid of negative charges from carboxylates. This is apparent on Figure 4, which illustrates the electrostatic potential on the protein surface in four complexes containing double-stranded DNA and respectively the lambda phage

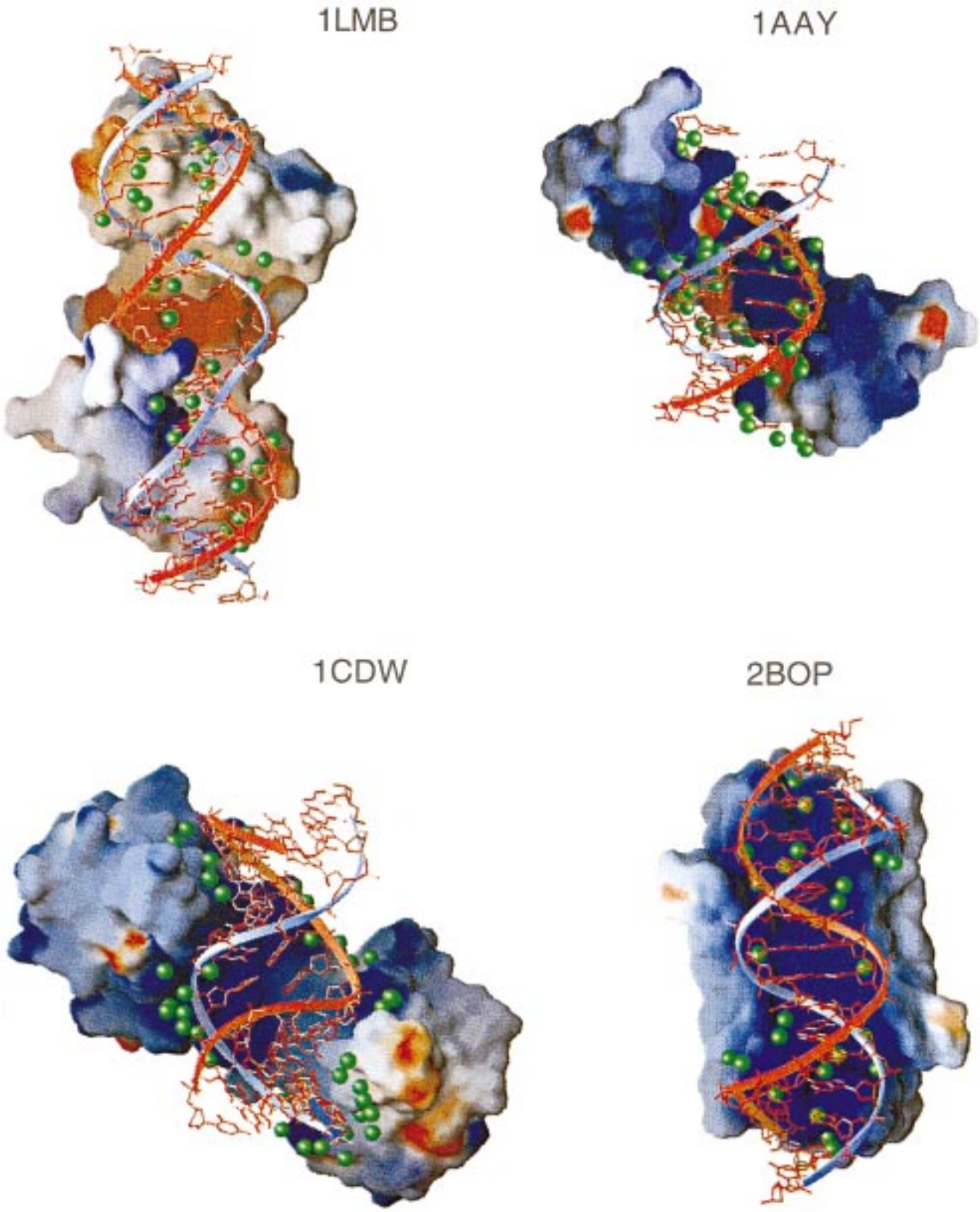

FIGURE 4: Examples of protein interaction with double-stranded DNA. The protein surface is colored according to its electrostatic potential, positive in blue and negative in red. Green spheres are interface water molecules: 1lmb, the dimeric lambda repressor; 1aay, three zinc fingers in the Zif268 transcription factor; 1cdw, human TBP; and 2bop, the dimeric papillomavirus E2 domain. Drawn with GRASP (*116*).

repressor (1lmb), the Zif 268 zinc fingers (1aay), human TBP (1cdw), and the papillomavirus E2 domain (2bop). In the last three proteins, the surface in contact with DNA is colored uniformly blue indicating a strong positive potential. The electrostatic potential is more contrasted on the surface of the lambda phage repressor. It has a red negative region near the 2-fold axis of the dimer at the center of the molecule, which is not in direct contact with DNA; positive charges concentrate in a N-terminal arm which wraps around the double helix. In the figure, the arm is represented only for

Table 5: Nucleotide/Amino Acid Composition of Interfaces[a]

residue type	DNA complexes		protein−protein complexes		oligomeric proteins	
	surface	interface	surface	interface	surface	interface
nucleotides						
G	23.8	23.4				
A	26.7	24.6				
C	22.5	20.5				
T	27.1	31.5				
amino acids						
Ala	3.4	3.4	4.0	2.6	5.9	4.1
Arg	12.1	23.8	8.9	10.1	8.4	9.9
Asn	5.3	6.3	6.2	5.5	5.2	4.6
Asp	6.4	1.6	7.1	5.2	7.8	4.8
Cys	0.4	0.8	0.7	1.5	0.4	0.8
Gln	5.5	5.1	6.0	4.2	5.4	3.5
Glu	12.3	2.5	9.8	6.1	10.3	4.1
Gly	3.3	3.6	4.5	4.6	4.8	4.2
His	2.9	3.8	1.9	3.6	3.5	4.5
Ile	2.8	2.8	2.4	4.2	2.2	4.6
Leu	5.1	2.4	4.1	5.5	3.8	10.5
Lys	16.5	17.5	11.8	6.7	14.9	5.4
Met	1.8	1.2	1.2	3.2	1.5	3.9
Phe	1.8	3.8	2.0	4.4	1.9	6.0
Pro	4.3	2.2	5.1	4.0	5.6	5.3
Ser	4.7	6.3	8.4	5.5	6.3	4.1
Thr	4.4	6.7	7.3	5.1	5.5	4.7
Trp	0.8	0.5	1.3	4.5	0.8	2.4
Tyr	3.4	3.4	3.2	9.1	2.7	5.4
Val	2.9	2.4	3.6	3.8	3.2	7.3
nonprotein groups	0.6	0.5	0.7	0.6		
charged/ hydrophobic[b]	3.3	3.6	2.9	1.3	3.3	0.7

[a] Compositions are given as percent contributions to the protein or DNA surface area. Data for protein−protein complexes are from ref *22*, for the accessible surface and the subunit interfaces in oligomeric proteins, from ref *21*. [b] Ratio of(Arg, Lys, Asp, Glu) to (Ile, Leu, Val, Phe, Met).

the bottom subunit of the dimer, because the crystal structure shows that it is disordered in the top subunit.

The abundance of positive charges on the protein side of the interfaces has its counterpart in the DNA surface where the negatively charged phosphate group dominates and there is no positive charge at all. The negatively charged component represents one-third of the accessible surface of double-stranded DNA and an even larger fraction (43%) of the surface in contact with proteins. Nucleobases contribute to the same extent (≈27%) to interfaces and to accessible surfaces, but sugars contribute less to the interfaces than to the accessible surface. The sugar surface is comparatively nonpolar, for there is no free hydroxyl in DNA. As a consequence, the nonpolar component from sugars and bases is smaller on the DNA surface in contact with proteins than on the accessible surface: 41% instead of 47%. Compared to the protein side, the DNA side of the interfaces is significantly more polar and has almost twice as much charged surface area. The excess is less striking when counting electric charges instead of areas. The average interface contains 15 negative charges from phosphates and 12 positive charges from lysines or arginines. Presumably, metal cations ensure electrostatic neutrality, but very few are located in electron density maps and reported in the PDB entries.

Amino Acid/Nucleotide Composition. Table 5 describes the average residue composition of the solvent accessible surface and of the interfaces in the 65 double-stranded DNA−protein complexes. The compositions are given as a percentage of the interface or accessible area for each type of nucleotide or amino acid. For comparison, average protein surface compositions are also cited for the solvent accessible surface and the subunit interfaces in samples of protein−protein complexes (*22*) and oligomeric proteins (*21*).

On the DNA side of the complexes, interfaces have a nucleotide composition similar to the surface that remains solvent accessible. Our sample has a 15% excess of A/T over G/C residues. The excess is also 15% among interface residues, which is to be expected for they represent two-thirds of the nucleotides present in the PDB entries we used. The contributions of the four nucleotides to the DNA accessible surface area show the same excess of A/T over G/C. However, this is no longer so at interfaces. Table 5 shows that T contributes 50% more area than C to the average interface. The 5-methyl group of thymine appears to be responsible for this larger contribution. It represents 8% of the interface area on the DNA side and is the chemical group that contributes the most area to the interface except for the phosphate. Its interactions, recently reviewed in ref *23*, include nonpolar contacts and CH···O hydrogen bonds.

On the protein side, the two basic amino acids, arginine and lysine, dominate interfaces: together they account for 41% of the area. Lysine is abundant on the protein surface, and only marginally more so in regions where it interacts with DNA. This differs from protein−protein interfaces and subunit interfaces in oligomeric proteins, which are depleted in lysine. Arginine is generally less abundant in proteins than lysine, but it is over-represented by a factor of at least two at interfaces with DNA, relative to either the accessible surface or the other types of interfaces. Another striking peculiarity of interfaces with DNA is the exclusion of the acidic residues: Asp and Glu together account for 4% of their area versus 19% for the accessible surface and 11% of protein−protein interfaces. The amino acid composition of the interfaces with DNA explains the presence of a large positively charged and a very small negatively charged component to their area. Concerning hydrophobic amino acids, the composition of the interfaces with DNA differs little from that of the surface that remains accessible; the more hydrophobic residues (Ile, Leu, Val, Phe, Met) contribute 13−14% to both. They contribute significantly more (22%) to protein−protein interfaces, and much more (32%) to subunit interfaces in oligomeric proteins (*21*−*22*, *24*).

For pairwise comparisons, a Euclidian distance can be computed in the 19-dimension space of amino acid compositions and expressed as

$$\Delta f^2 = (1/19) \sum_i (f_i - f'_i)^2$$

where f_i and f'_i are the percent areas contributed by residue type i to the two surfaces. Δf values for some of the compositions in Table 4 are shown in Figure 5. The distance between the average solvent accessible surfaces of the two types of complexes and the oligomeric proteins is 1.4−1.8%, which may be taken as a background value. That of 3.9% between the surface and the interior of a sample of oligomeric proteins represents the large difference in the amino acid composition of the protein surface that is buried upon folding and that which remains solvent accessible. Subunit interfaces

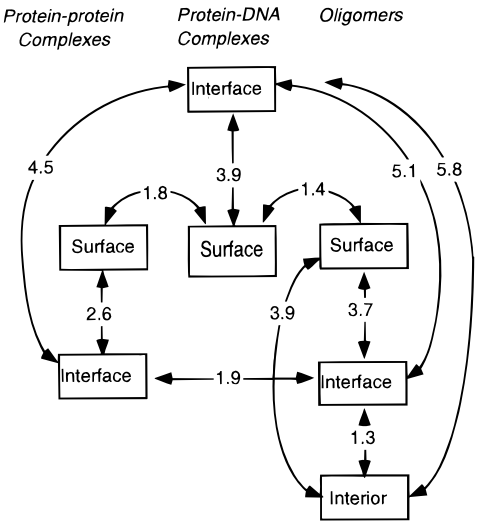

FIGURE 5: Distances between amino acid compositions. Euclidian distances (in percents) are calculated from data in Table 5 as explained in the text.

and protein−protein interfaces are between the protein surface and its interior. Interfaces with DNA are much more polar than these two types of interfaces, and even more polar than the average protein surface. Their composition is highly specific, as shown by the large distances to the accessible surface of the complexes and to other types of buried surfaces.

Polar Interactions. Hydrogen bonds. As interfaces bury a large amount of polar surface, buried polar groups form many polar interactions, which bridge the protein and the nucleic acid. A study of these interactions has been reported on a smaller sample (*25*). In Table 2 above, we quoted the number N_{hb} of protein−nucleic acid hydrogen bonds that were detected with HBPLUS (*26*) using the default geometric parameters of the program. The geometry was good, with 98% of the donor−acceptor distances shorter than 3.4 Å and 67% shorter than 3.0 Å. Table 6 summarizes these data and details the types of bonds that are found in complexes with double-stranded DNA. No similar detail is given for single-stranded DNA and RNA due to the small number of complexes. The number of hydrogen bonds per complex varies widely, from 2 between thrombin and the single-stranded DNA aptamer (1hut) to 58 between glutaminyl−tRNA synthetase and the cognate tRNA (1gtr). On average, there are 22 hydrogen bonds in a complex with double-stranded DNA. With single-stranded DNA and RNA, the corresponding number is quoted in Table 6, but the statistics are poor.

N_{hb} generally increases with the size of the interface (Figure 6). However, the correlation is mediocre, partly because hydrogen bond detection is sensitive to the quality of the X-ray structures. In structures of complexes with DNA having 2.4 Å or better resolution, there is an average of one hydrogen bond per 125 Å^2 of interface area and the correlation coefficient between N_{hb} and B is 0.73. This value should be retained instead of the one listed in Table 6 for all complexes, because the comparatively fewer hydrogen bonds detected in structures with poorer resolution are likely artifacts. It implies that a recognition module with a 1600 ± 400 Å^2 interface area contains about 13 hydrogen bonds.

Table 6: Hydrogen Bonds at Protein−Nucleic Acid Interfaces

type	DS-DNA	SS-DNA	RNA
average number per complex[a]			
H-bonds	21.8 (8.0)	9.5	24.5
salt bridges	5.2		
average B per H-bond (Å^2)[b]			
high-resolution structures	126	125	
all complexes	146	181	127
chemical group			
(fraction of total H-bonds)			
nucleic acid			
phosphate	0.60	0.47	0.25
sugar	0.06	0.18	0.31
base	0.34	0.34	0.44
G	0.16		
A	0.07		
C	0.07		
T	0.04		
Protein			
Main/side chain acceptor O	0.10	0.10	0.28
Main chain N donor	0.18	0.13	0.15
Side chain donors	0.73	0.76	0.56
Arg,Lys	0.41		
other N (Asn, Gln, His, Trp)	0.14		
-OH (Ser, Thr, Tyr)	0.17		
-SH (Cys)	0.01		

[a] The value in parentheses is the standard deviation, quoted only for double-stranded DNA complexes. For these complexes, the total number of H-bonds is 1416. It is 38 in 4 complexes with single-stranded DNA, and 147 in 6 complexes with RNA. Hydrogen bonds include salt bridges. [b] Interface area divided by the number of H-bonds. The high-resolution structures are 25 double-stranded DNA, two single-stranded DNA and one RNA complexes with resolution 2.4 Å or better.

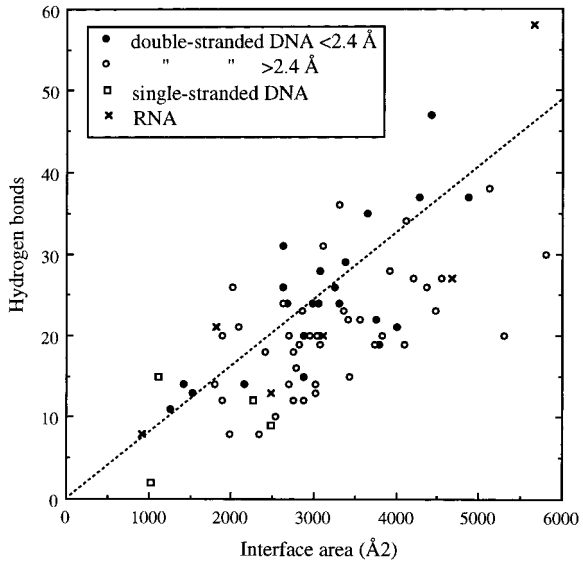

FIGURE 6: Correlation between the number of hydrogen bonds and the interface area. For structures with resolution 2.4 Å or better (black dots), the correlation coefficient between B and N_{hb} is 0.73. The dashed line corresponds to one bond per 125 Å^2.

Ninety percent of the protein-nucleic acid hydrogen bonds have the donor group on the protein and the acceptor group on the DNA or RNA. This results directly from the chemical nature of the two types of macromolecules. In double-stranded DNA, the phosphate group is involved in 60% of hydrogen bonds and the deoxyribose sugar 3′, 4′, and 5′ oxygens in 6%, all as acceptors. The nucleobases provide the remaining 34%, also mostly as acceptors. On average, a complex with double-stranded DNA contains 13 bonds to

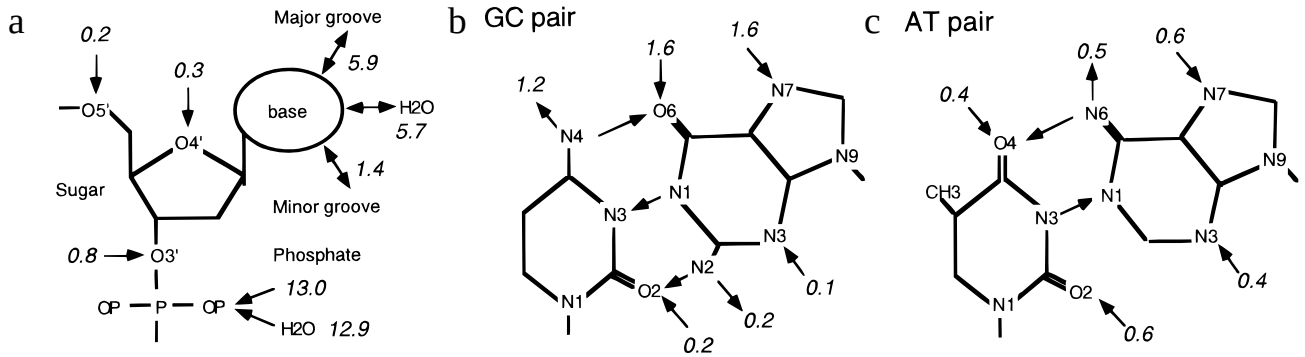

FIGURE 7: Polar interactions to double-stranded DNA. The arrows represent protein−DNA hydrogen bonds and point to the acceptor atom. Numbers in italics are the average number of hydrogen bonds of each type per complex, or of water bridges per complex at 2.4 Å or better resolution: (a) distribution of polar interactions between the phosphate, the sugar oxygens, and the base; (b) hydrogen bonds to a GC pair; and (c) hydrogen bonds to a AT pair.

the phosphates, 1.4 to the sugar, and 7.3 to the bases (Figure 7a). The distribution of hydrogen bonds is about the same in single-stranded DNA complexes, but not in RNA complexes. There, bonds to the sugar are much more abundant and the free 2′-hydroxyl plays a major part, second only to the phosphate. We find the 2′-hydroxyl of ribose to be involved in 20% of the interface hydrogen bonds, equally often as a donor and as an acceptor (data not shown).

On the protein side, the neutral main chain NH group and the charged side chain groups of lysine and arginine are the major hydrogen bond donors. Together, they account for 60% of the hydrogen bonds to DNA and 65% of those to the phosphates. The two positively charged groups can form salt bridges with the phosphates. There is an average of ≈5 salt bridges per complex, representing 40% of all phosphate hydrogen bonds. Neutral protein donor groups give the remainder. In addition to the main chain NH, these groups are side chain hydroxyls, amide, and indole or imidazole groups. Among the few hydrogen bond acceptors found on the protein side, the main chain carbonyl is the one most often implicated in bonds with DNA groups.

Direct Recognition of DNA Bases. In double-stranded DNA, direct base recognition by protein groups can occur in the major or the minor groove (Figure 7a). Direct base recognition also takes place in single-stranded DNA or RNA, but statistics are insufficient to derive general rules. Polar interactions in the major groove concern groups in positions 6 and 7 of purines or position 4 of pyrimidines. They represent 80% of the hydrogen bonds to the bases and an average of ≈6 bonds per complex. Minor groove interactions with positions 2 and 3 of purines and position 2 of pyrimidines are four times less abundant. The GC and AT pairs behave very differently in these bonds: GC makes three times as many major groove hydrogen bonds as AT; and AT makes twice as many as GC in the minor groove.

Guanine accounts for almost half of the bonds to bases, essentially through its O6 and N7 atoms, both hydrogen bond acceptors located in the major groove (Figure 7b). In 80% of the cases, the donor group is from a lysine or an arginine side chain. There is an average of 2.5 hydrogen bonds per complex relating a Lys/Arg side chain to a guanine O6/N7 atom. In over half of the cases, it is the same residue that donates hydrogen bonds to both acceptor atoms. Direct recognition of guanine in the major groove by a Lys/Arg side chain occurs at least once in 45 of the 65 complexes. Due to the sequence specificity or the type of oligonucleotide that was used in crystallization, it is not found in complexes with TBP, DNA polymerases, and several of the nucleases, but it is present in 80% of the remaining complexes. In the major groove, guanine has a strong electronegative character (*27)* and its complementarity to the guanidinium group of arginine has long been noted (*28*), but the amino group of lysine also often interacts simultaneously with both O6 and N7. For the complete recognition of the GC pair in the major groove, the N4 amino group of cytosine is the most significant hydrogen bond donor in DNA and it receives an average of 1.2 hydrogen bonds per complex. Its partner on the protein can be a main chain carbonyl or a side chain carboxylate oxygen, in equivalent numbers.

Recognition of adenine in the major groove is performed almost exclusively by Asn/Gln side chain amide groups. They represent 90% of the protein groups interacting with the N6 donor and/or the N7 acceptor group (Figure 7c). In 70% of the cases, the same residue interacts with both groups, illustrating another well-established case of side chain−nucleobase complementarity. On average, we find an average of 1.0 Asn/Gln−adenine hydrogen bond per complex. However, adenine recognition by Asn/Gln occurs in only 17 of the 65 complexes and its role is clearly less general than guanine recognition by Lys/Arg. Also, bonds to O4 of thymidine are much less common than to N4 of cytosine.

In the minor groove, base recognition involves N3 in purines and O2 in pyrimidines. They accept a hydrogen bond from a protein donor group with no obvious preference for any type of donor. On average, there are 1.2 hydrogen bonds of this type per complex. In rare cases, the N2 amino group of guanine donates a hydrogen bond, usually to a main chain carbonyl.

Water Molecules and Metal Ions. Solvent molecules are commonly found at protein-nucleic acid interfaces. Their importance, first emphasized in the tryptophan repressor−operator complex (*29*), is obvious in many other structures. Because the identification of water in the electron density map may be ambiguous at lower resolution, we restricted our analysis to 28 structures with 2.4 Å or better resolution. A solvent molecule was defined as being at the interface if it was within 3.5 Å from atoms of both the protein and the nucleic acid. With very few exceptions, interface solvent molecules hydrogen bond to both macromolecules. Table 2 gives their number N_{wat} in the 28 structures. Its range is 2−43, with an average of 21, close to the average number of direct hydrogen bonds. Per complex, there is an average

of 13 water molecules bridging the DNA phosphates to the protein and 5.7 bridging the bases.

On average, there is one water molecule per 150 Å^2 and the correlation coefficient between N_{wat} and B is 0.66. Crystallographers have different practices in reporting water molecules, which partly explains the mediocre correlation. Yet, the data suffice to show that, in most complexes, there are at least as many water-mediated polar interactions between the protein and nucleic acid as there are direct hydrogen bonds. The location of interface water molecules was illustrated in Figure 4 above for the lambda phage repressor (1lmb), the Zif 268 zinc fingers (1aay), the human TBP (1cdw) and the papillomavirus E2 domain (2bop). Water, drawn as green spheres, tends to follow the phosphate backbone of DNA. In the TBP complex, water is excluded from the center of the interface where the sugar moiety of the DNA is in contact with a rather hydrophobic region of the protein surface. It forms an almost continuous ring around it. In the other three complexes, it is distributed more uniformly and bridges the protein to the DNA bases as well as the phosphates.

Metal ions are reported in a small number of structures, essentially as Mg^{2+} or other divalent cations. They are not involved in protein−nucleic acid interfaces with the important exception of the catalytic sites of polymerases and nucleases. Other cations must be present to maintain electric neutrality, but they either have no specific site or cannot be distinguished from water molecules in the electron density map.

Atomic Packing and Geometric Complementarity. The packing density of atoms at an interface is an index of the shape complementarity between the two surfaces in contact. It is related to the volumes occupied by atoms, which can be estimated by constructing a Voronoi polyhedron around each atom and calculating its volume (*30*). The packing density for the protein component of the protein-nucleic acid interfaces was determined by calculating the volumes of individual interface atoms, summing the values to give a total volume V, and comparing V to a reference value V_o. To derive V_o, we used the mean volumes that the corresponding atoms occupy in the protein interiors. A V/V_o ratio larger than unity means that the packing density at interfaces is lower than inside proteins, and a ratio less than unity means a higher packing density. Due to the lack of a set of reference volumes, similar calculations could not be performed for the nucleic acid component. For the sake of accuracy, we restricted our analysis to complexes with resolutions of 2.4 Å or better, which numbered a total of 28.

Volumes of atoms buried in the interfaces were calculated using SURVOL (*13*), which implements a variant of the Voronoi procedure of Richards (*30*). The faces of the Voronoi polyhedra were defined by radical planes positioned in proportion to the van der Waals radii. The same procedure was used previously to determine the mean volumes of atoms buried inside globular proteins (*15*). Harpaz et al. (*31*) have shown that these volumes are 5% smaller than the volumes of equivalent groups in small molecule crystals of amino acids, indicating that the protein interior is on the average more tightly packed than small molecule crystals, despite the occasional presence of internal cavities. As small molecule crystals are usually considered to be close-packed, volumes observed inside proteins are a good reference for optimal atomic packing.

A Voronoi polyhedron cannot be built around an atom unless it is fully surrounded by other atoms. This implies that the calculation can only be carried out on buried atoms. On average, only 28% of all interface protein atoms were found to be buried. We calculated their volumes in each of the higher-resolution complexes in our sample. Only two contain single-stranded DNA and one, RNA. In Table 7, we list the number of interface atoms, the fraction of the interface atoms that are buried, and the values of the V/V_o volume ratio observed at each interface. A histogram of the V/V_o values is shown in Figure 8. The mean is 1.02 which indicates that, on average, the packing density of protein atoms forming interfaces with DNA is very similar to that within the protein interior. However, individual values spread over a rather wide range, 0.94−1.10, and the volume measurement applies only to a minority of the interface atoms. On average, 50 interface protein atoms per complex were fully buried, and in half of the complexes, they were fewer than 40. The spread of the volume ratio may result in part from fluctuations of V due to this small number and not express real differences in packing density.

This weakness of our procedure, which cannot handle the 72% interface atoms that have residual accessibility in the complex, could be overcome in part by taking solvent into account. The proportion of interface protein atoms with zero accessibility doubled to 57% when reported solvent positions were included in computing the Voronoi polyhedra. Fifty protein atoms were added in the calculation for the average complex. These atoms are surrounded by a combination of protein atoms, nucleic acid atoms, and the crystallographically determined solvent molecules. No volume was calculated for the solvent molecules themselves. The new volume ratio is quoted as V'/V_o in Table 7. Its mean is 1.01 and the spread narrower than for V/V_o: all but one value are in the range 0.97−1.04 (Figure 8), which indicates that the protein side of the interfaces is as close-packed as the protein interior. In the papillomavirus E2 domain (2bop), three-quarters of interface atoms could be taken into account, and the volume ratio was 1.00. In this complex, crystallographic water positions occur both at the periphery and inside the interface (Figure 4), and they play an important part in determining the shape complementarity as indicated by the volume ratio. In several other complexes, close packing can be established for most of the interface atoms on the protein side, though no conclusion can be drawn at present on the DNA side.

Interface atoms in the p53 oncogene protein (1tsr) have large V/V_o and V'/V_o volume ratios. The second value, 1.08, stands out from the rest. It applies to only 25 atoms, due in part to the small size of the interface and in part to the relatively low fraction of atoms with zero accessibility. Nevertheless, it confirms that the p53 interface has features which are unique in our sample, with an interface area of only 1250 Å^2, very few buried atoms, and comparatively poor packing at the interface. In the other copy of the same protein fragment that is in contact with DNA in the crystal structure, these features are even more pronounced.

Conformation Changes. Conformation changes can be evaluated by comparing the structure of the components in a complex and in free state in cases where the latter has been independently determined. For the protein component

Table 7: Packing Density of Protein Atoms at Nucleic Acid Interfaces[a]

code	protein		interface atoms	V/V_o	buried atoms (%)	V'/V_o	buried (with water) atoms (%)
double-stranded DNA (*25*)							
1bpy	DNA polymerase β		188	1.03	35	1.00	66
1t7p	Phage T7		254	1.05	32	1.04	53
1bhm	*Bam*HI		251	1.03	27	1.01	67
1dnk	DNase I		93	1.02	24	1.02	44
1rvc	*Eco*RV		298	1.03	26	1.02	70
1hcr	*Hin* recombinase		145	0.98	36	1.01	39
1tc3	*C. elegans* transposase		103	0.98	37	1.00	48
1ign	RAP1 telomere binding		233	0.97	30	0.97	53
1lmb	lambda repressor		163	1.06	28	1.02	54
1tro	Trp repressor		152	1.10	24	1.04	56
1trr	Trp repressor, half-site		167	1.06	30	1.01	49
1fjl	paired dimer		184	0.99	19	0.98	72
1pue	Pu1-ETS domain		103	1.02	28	1.00	50
2dgc	GCN4, ATF site		120	1.07	18	1.01	38
1aay	Zif 268		146	1.01	15	1.00	65
1mey	designed		138	1.09	10	1.00	59
1hcq	estrogen receptor		141	1.10	27	1.04	52
1lat	glucocorticoid		75	1.07	31	1.00	73
2nll	retinoid receptor		175	1.04	24	1.02	51
1ais	TBP-TFIIB		216	1.01	40	1.03	62
1cdw	TBP, human		152	1.02	51	1.02	73
1a3q	NFκ-B p52		182	0.99	16	1.03	59
1nfk	NFκ-B p50		203	0.95	16	1.02	48
1tsr	p53 core		63	1.10	25	1.08	40
2bop	E2 domain		166	1.03	25	1.00	76
single-stranded DNA (*2*)							
1brn	barnase		79	0.98	42	0.99	77
1jmc	RPA70		136	0.94	32	0.97	38
RNA complex (*1*)							
1urn	spliceosome U1A		98	0.95	37	0.95	67
all double-stranded DNA		average	164	1.03	27	1.02	56
complexes		sd	57	0.04	9	0.02	12
all 28 complexes		average	158	1.02	28	1.01	57
		sd	58	0.05	9	0.02	13

[a] V is the sum of the volumes of the protein atoms buried at interfaces with nucleic acid, V', the same value calculated in the presence of crystallographic water molecules, V_o, a reference volume for atoms buried inside proteins. The total number of protein atoms at the interfaces and the fraction that are buried and included in the volume calculation are quoted.

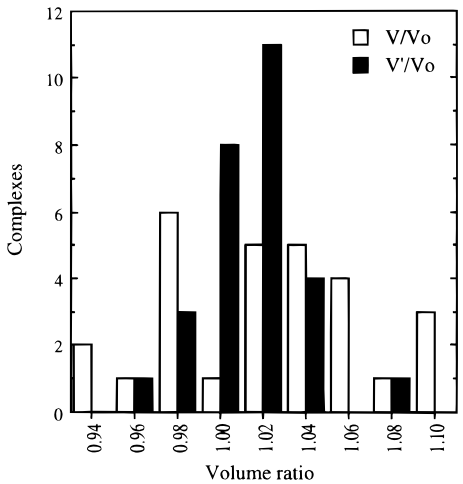

FIGURE 8: Histogram of packing densities for protein atoms at interfaces with nucleic acids. Distribution of the values of the V/V_o and V'/V_o ratios cited in Table 7 for protein atoms at the interfaces of 28 complexes with resolutions of 2.4 Å or better. V is the sum of the Voronoi volumes of buried interface atoms; V' is the corresponding value for atoms with zero accessibility in the presence of crystallographic water molecules; and V_o is a reference volume observed for atoms buried inside proteins.

of 24 complexes, the structure of the free molecule was available in the PDB. All equivalent backbone atoms in the two structures were superimposed by least squares. Residues undergoing backbone movements of less than 1 Å were taken to form an invariant core, and superposition was repeated on the basis of these residues only. Conformation changes in the protein were then characterized by the root-mean-square value of the residual distance (RMSD) between equivalent backbone atoms in all interface residues in the second superposition. For double-stranded DNA, we took as a reference a canonical 40-bp B-DNA model (Srinivasan, A. R. and Olson, W. K., personal communication) and performed least-squares superposition on phosphorus positions. We made no attempt to model single-stranded DNA. The RMS phosphorus distance was used as a measure of the conformation change. Albeit a crude estimate, RMSD values from least-squares superposition are an easy way of detecting movements. They suffice for our purpose, for conformation changes affecting the protein and the nucleic acid have been extensively described for individual systems in the publications listed in Table 1.

Changes in the Protein. In the protein component, several types of conformation changes were distinguished. Their occurrence in 24 complexes is reported in Table 8. (a) Disorder-to-order transitions occur as disordered parts of the polypeptide chain in free state become ordered in the complex. The corresponding residues are usually missing from the PDB file. The abundance of disorder-to-order

Table 8: Conformation Changes in the Protein[a]

code	protein	percent residues		RMSD (Å)		ΔB[b]	Type of change			
		core	interface	core	interface	(Å^2)	order	quaternary	domain	tertiary
enzymes										
1kln-1dpi	Klenow fragment	64	8	0.5	1.5	N/A	+	−	+	+
1tau-1taq	Taq	92	6	0.4	1.4	3.6	−	−	−	+
1bhm-1bam	*Bam*H1	50	19	0.5	5.2 2.0	5.0	+	+	+	+
1dnk-3dni	DNase I	99	13	0.3	0.6	−0.7	−	−	−	−
1fok-2fok	*Fok*1	52	11	0.7	1.4	15.	+	−	−	+
1rvc-1rve	*Eco*RV	53	18	0.5	1.5 2.0	8.	+	+	+	+
1pvi-1pvu	*Pvu*II	16	26	0.7	2.6	6.7	−	+	+	+
1vas-2end	endonuclease V	96	33	0.4	0.6	0.4	−	−	−	−
1mht-1hmy	*Hha*I methyl transferase	49	17	0.4	7.9	9.	−	−	+	+
prokaryotic transcription factors										
1cma-1cmb	Met J repressor	65	14	0.5	2.6	24.	−	−	−	+
1lmb-1lpr	lambda repressor	75	27	0.6	0.9 1.2	N/A	+	−	+	+
1par-1baz	Arc R	70	27	0.5	2.4 0.6	2.3	+	−	+	+
1per-1r69[c]	434 repressor	97	35	0.5	0.6	2.2	+	+	−	−
1ruo-3gap	CAP	65	11	0.6	1.3 4.3	5.0	−	+	+	+
1tro-3wrp	Trp repressor	57	23	0.5	1.9	N/A	−	−	+	+
3cro-2cro	434 Cro repressor	92	39	0.5	0.5	4.	−	+	−	−
eukaryotic transcription factors										
1hdd-1enh*	engrailed	83	37	0.4	2.4	6.7	+	+	+	+
1ais-1pcz	TBP, *Pyrococcus*	57	28	0.7	1.1	−2.	−	−	−	+
1cdw-1tbp	TBP, human	57	26	0.7	1.3	−0.	−	−	−	+
1vol-1vok	TBP, *Arabidopsis*	78	25	0.6	0.9	0.6	−	−	−	+
1tsr-1ycs	p53 core	90	9	0.5	1.4	3.0	−	−	−	+
RNA complexes										
1ser-1ses	Seryl-tRNA synthetase	92	2	0.6	1.2 4.5	21.	−	−	+	+
1ttt-1eft	EF-Tu	91	14	0.5	0.9	2.7	−	−	−	+
1urn-1nrc	U1A protein	90	29	0.4	0.9	12.7	−	−	−	+

[a] PDB codes refer to the complex and the free protein. The percent fractions of all protein residues that are included in the core and in the interface are quoted with the corresponding RMSD after superposition of the core. Conformational changes observed in the protein are categorized into "order" (disorder-to-order transitions), "quaternary" (dimerization or whole subunit movements in dimers), "domain" (rigid-body domain movements), and "tertiary" movements. Their presence in a complex is marked with a + sign. [b] ΔB is the double difference defined in the text for the temperature factors of main chain atoms. Positive values indicate that interface atoms become less mobile in the complex. N/A: temperature factors were not available for comparison. [c] The protein binds DNA as a dimer, but the free protein is a monomer.

transitions is likely to be underestimated in our data, for many proteins that undergo such transitions do not yield useful crystals or NMR spectra in free state. (b) Quaternary structure changes take place when a monomeric protein forms a dimer, or when subunits of a dimer move relative to each other upon DNA binding. In the second case, the change was identified by performing the least-squares superposition separately on the two subunits. (c) Within a monomeric protein or a subunit, domains may undergo rigid-body movements. These were identified in the same way as subunit movements. (d) Other changes within a subunit are listed as tertiary changes and characterized by the RMSD obtained by superposing main chain atoms in the subunit. (e) In addition, side chain rotations occur in all complexes.

Table 8 reports disorder-to-order transitions in 8 of the 24 complexes, quaternary structures changes in 7, domain movements in 11, and tertiary changes in 20. Within subunits, a RMSD of 0.4−0.7 Å is observed for the invariant core. The range is typical of differences observed between two crystal structures of the same protein (*32*). In seven of the complexes, the main chain of interface residues has an RMSD less than 1 Å. Thus, the tertiary structure is essentially unchanged though side chains may still undergo large movements. In the complexes with the phage 434 and Cro repressors (1per, 3cro), the quaternary structure changes, but in the other five, the change is restricted to short loop movements and the protein associates with DNA or tRNA approximately as a rigid body.

In twelve other complexes, the RMSD is in the range 1−3 Å. In that range, local changes occur within subunits, involving for example the movement of a large loop or a whole α-helix, and they are often coupled with domain movements or changes in the quaternary structure. Figure 9 illustrates such medium size conformational changes for the human TBP (1cdw/1tbp), where backbone adjustments affect residues at the interface with DNA, and particularly the loop between residues 185−195.

The remaining four complexes have an RMSD of 4−8 Å for interface atoms. In these complexes, the region of the protein surface that is in contact with DNA is largely remodeled. In three, the protein is a dimer and the RMSD is significantly larger in one subunit than in the other, due to an asymmetry in the protein or the mode of binding. The *Bam*HI endonuclease (1bhm/1bam, Figure 10) is a good example of such remodeling, which in this protein is achieved through conformational changes at the quaternary and tertiary as well as secondary structure levels (*33*).

FIGURE 9: Conformational changes for human TATA binding proteins. The DNA-bound complex (1cdw) is shown in dark blue and the free form of the protein (1tbp) displayed in cyan. Interface residues are colored in red. The protein undergoes local changes compatible with a 1.3 Å RMSD. The DNA, traced in green and yellow, is strongly distorted. Drawn with MOLSCRIPT (*117*).

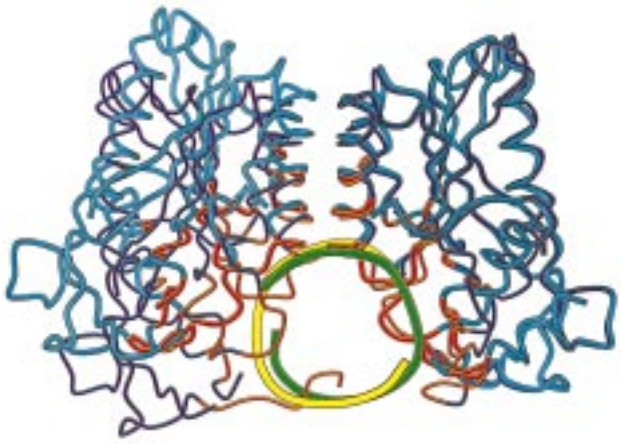

FIGURE 10: Conformational changes for *Bam*HI endonuclease. The figure shows the backbone of the dimeric *Bam*HI endonuclease after superposing one subunit in the complex (1bhm) and in the free form (1bam). When DNA binds, conformational changes take place at the level of the quaternary, tertiary, and secondary structures. The complex is shown in blue and the free form in cyan. Interface residues are displayed in red. The protein undergoes a quaternary structure change, seen here as a global movement of the subunit on the left. Additional local movements of large amplitude can be observed in both subunits. The DNA, traced in green and yellow, deviates moderately from the B form (RMSD 2.1 Å). Drawn with MOLSCRIPT (*117*).

In addition to conformation changes, the interaction with DNA or RNA tends to reduce the mobility of the protein atoms. This can be deduced from calculating the average crystallographic temperature factors of main chain atoms in the complex and the free protein for the interface residues and comparing the difference to that observed for the whole protein (temperature factors, B, should not be confused with interface areas noted, *B*, in this paper). Table 8 reports values of the double difference:

$$\Delta B = \langle B_{free} - B_{complex} \rangle_{interface} - \langle B_{free} - B_{complex} \rangle_{whole\ protein}$$

B_{free} and $B_{complex}$ refer to the crystallographic temperature factors in the free and complexed proteins, respectively, and the differences are averaged (brackets) over main chain atoms of either the interface or the entire protein. In 16 out of the 21 complexes for which temperature factors were available, ΔB was positive with values ranging from 2 to 26 Å^2. This suggests that, on average, the mobility of protein interface atoms decreases more upon DNA binding than for core atoms. Also, a large differential reduction in mobility often accompanies large conformational changes. In the remaining five cases, ΔB is marginally negative or 0, indicating no change in mobility. Though the accuracy of the crystallographic temperature factors is sensitive to resolution, these trends are quite clear.

Changes in DNA and tRNA. Table 9 reports the number of phosphorus atoms superimposed to the B-DNA model in each of the 65 double-stranded DNA complexes, and the RMSD to the model. Terminal phosphates and unpaired nucleotides were excluded from the superposition in order to limit the effect of having free ends. Thus, the number of nucleotide residues is less than in Table 1, but it exceeds the number of interface nucleotides in Table 2 by 23% as some of the atoms used in the superposition belong to nucleotides outside the interface. Observed RMSD to canonical B-DNA vary between 1.2 and 22 Å. The lower limit is close to the RMSD between the model and the X-ray structure of a 12-bp fragment of free B-DNA (1dn9). It may therefore be taken as a baseline value reflecting the natural variability of B-DNA.

Five of the DNA fragments in the protein complexes show a RMSD of less than or equal to 1.5 Å and may be considered as having a standard B-DNA conformation. They are bound to the *Hin* recombinase (1hcr), the yeast pyrimidine pathway regulator protein (1pyi), the estrogen receptor (1hcq) and NFκ-B (1a3q, 1nfk). Thirty-two DNA fragments, or half of the sample, have a RMSD between 1.5 and 3 Å. They undergo deformations of limited amplitude and of various types. The most frequent are bending or unwinding of the double helix, and the widening of the major groove where most of the interactions with the protein take place. In the lambda repressor complex, for instance, the DNA fragment has been described as essentially B-form, with a wide major groove and significant variations from B-DNA in twist and other helical parameters (*34*). These changes are compatible with the 2.6 Å RMSD from the canonical B-form quoted in Table 9.

In the remaining 29 complexes, or 45% of the sample, DNA undergoes major changes leading to a RMSD of more than 3 Å. In seven, it is over 8 Å. These are the four complexes with TBP, and the complexes with $\gamma\delta$-resolvase (1gdt), the catabolite activator protein (CAP, 1ruo), and the integration host factor (1ihf). The double helix is dramatically perturbed in these complexes. When bound to the TBP, it is unwound and bent at a right angle (*35*, *36*); in the complex with $\gamma\delta$-resolvase, it is sharply bent by 60° at the point where recombination would occur (*20*). CAP bends DNA at two points, to a total of 90° (*37*); and the bending angle reaches 160° in the complex with the integration host factor, around which the double helix completes a U-turn (*62*). These distortions play a major functional role in each case (*38*).

A comparison of Tables 2 and 9 shows that all the DNA fragments in the seven complexes with $B < 2000$ Å^2 have a RMSD in the range 1.2−3.1 Å. Therefore, they undergo only

Table 9: Distortion of the DNA Double Helix and of tRNA in Complexes

code	protein	nucleotides[a]	RMSD[b] (Å)	code	protein	nucleotides[a]	RMSD[b] (Å)
DNA only				1pdn	paired	24	2.2
1dn9	B-DNA crystal	22	1.3	1yrn	Mat-A2/Mat-α2	36	4.5
DNA polymerases				leucine zippers, bHLH and others			
1bpy	DNA polymerase *β*	18	2.3	1an2	Max	42	3.3
1kln	Klenow fragment	18	1.7	1an4	USF	38	2.9
1tau	Taq, *T. aquaticus*	14	2.4	1fos	c-Fos/c-Jun dimer	36	2.0
1t7p	phage T7 polymerase	20	3.4	1hlo	Max intact dimer	20	2.8
endonucleases				1mdy	Myo D	26	2.2
1bhm	*Bam*HI	20	2.1	1pue	Pu1-ETS domain	28	3.7
1dnk	DNase I	12	1.6	1srs	serum response factor	34	5.1
1eri	*Eco*RI	22	3.2	1ysa	GCN4	36	2.4
1fok	*Fok*1	36	2.3	2dgc	GCN4, ATF site	34	2.8
1pvi	*Pvu*II	22	4.3	zinc fingers			
1rvc	*Eco*RV	18	4.8	1aay	Zif 268	18	2.6
1vas	phage T4 endonuclease V	22	6.6	1mey	designed	22	2.9
other enzymes				1ubd	YY1	38	2.9
1dct	*Hae*III cytosine methylase	32	4.6	2drp	Tramtrack	34	3.1
1gdt	*γδ*-resolvase	66	9.2	2gli	GLI1[b]	38	5.4
1hcr	*Hin* recombinase	24	1.3	other zinc modules			
1mht	*Hha*I methyl transferase[a]			1d66	Gal4	36	2.4
1tc3	*C. elegans* transposase	38	5.5	1glu	glucocorticoid receptor	34	2.9
miscellaneous				1hcq	estrogen receptor	32	1.5
1ecr	replication terminator	26	4.6	1lat	glucocorticoid (noncognate)	34	2.5
1ign	RAP1 telomere binding	34	3.2	1pyi	pyrimidine pathway regulator	26	1.2
1ihf	integration host factor	64	21.7	1zme	proline utilization PUT3	30	4.3
prokaryotic transcription factors				2nll	retinoid receptor	34	1.7
1cma	Met J repressor	16	2.3	TATA box binding protein			
1lmb	lambda repressor	36	2.6	1ais	TBP-TFIIB, *Pyrococcus*	28	13.1
1par	Arc R	40	4.3	1cdw	TBP, human	30	9.9
1per	434 repressor	36	2.7	1vol	TBP-TFIIB, *Arabidopsis*	30	9.7
1ruo	CAP	58	9.8	1ytf	TBP-TFIIA, yeast	30	8.6
1tro	Trp repressor	34	2.8	others			
1trr	Trp repressor, half-site	30	3.3	1a3q	NF*κ*-B p52, human	20	1.5
1wet	purine repressor	30	4.8	1nfk	NF*κ*-B p50, mouse	18	1.5
3cro	434 Cro repressor	36	2.5	1svc	NF*κ*-B p50, human	36	5.6
eukaryotic transcription factors				1tsr	p53 core	38	2.7
homeodomains				1xbr	T-domain	46	3.8
1apl	Mat-α2	38	2.7	2bop	E2 domain	30	3.3
1fjl	paired dimer	24	2.7	RNA[c]			
1hdd	engrailed	38	2.7	1asy-2tra	Asp-tRNA, yeast	73	5.1
1oct	Oct-1 POU	26	2.0	1ttt-4tna	Phe-tRNA, yeast	76	2.5

[a] 3′- and 5′-terminal and hanging nucleotides were not considered. [b] RMS residual distance of phosphorus positions after least-squares superposition onto a canonical B-DNA model. [c] Least-squares superposition of all phosphorus atoms in bound and free tRNA.

deformations of limited amplitude. At the other end of the scale, 12 out of 13 DNA fragments in complexes with $B >$ 4000 Å^2 have a RMSD larger than 3 Å. Thus, there is some correlation between the amplitude of the distortion away from the B-form measured by the RMS distance and the extent of the contact with the protein measured by the interface area. However, the linear correlation coefficient is only 0.5, and there are some fairly large interfaces where B-DNA is essentially unperturbed, with NF*κ*-B for instance.

Table 9 also reports changes observed in the tRNA component of the complexes with yeast aspartyl-tRNA synthetase (1asy) and the EF-Tu elongation factor (1ttt). The RMSD values are calculated on the whole tRNA, and the free tRNA structure is taken as the reference. In both complexes, the tRNA undergoes a significant distortion leading to a RMSD of 2.5 Å or more.

DISCUSSION

The present study was aimed at uncovering rules that apply across a variety of systems, and are not specific to one type of protein structure or interaction. It was inspired by a similar study of protein−protein recognition in protein−protein complexes of known structure (*39*). This study was recently extended on a sample of 75 protein−protein complexes (*22*), and its conclusions will be used for comparison.

An early result of the study of protein−protein recognition concerned the extent of the contact between macromolecules measured by the interface area B: there is a minimum size of the interface for stable association. The recent study shows that B is larger than 1250 Å^2 in all stable protein−protein complexes, and near 1150 Å^2 for short-lived complexes, between two redox proteins for instance. The present data on protein-nucleic acid recognition suggest that the minimum of B is the same ($\approx$1200 Å^2) in this process. While protein−nucleic acid interfaces are generally larger than protein−protein interfaces, both types of interfaces span the same range of areas, from about 1150 Å^2 to about 5000 Å. We suggest that the lower limit is set by physical requirements for stability, but not the upper limit, which is likely to be exceeded in the future.

Despite the spread of interface areas that is observed, the large majority (70%) of protein−protein complexes has an

interface with B in the range 1600 ± 400 Å^2, which is termed standard size in Lo Conte et al. (*22*). Standard-size interfaces occur in most protease−inhibitor and antigen−antibody complexes, two systems that are well-represented in the PDB. In contrast, most protein−nucleic complexes have interface areas above the range defining standard protein−protein interfaces. However, we have shown that complexes of double-stranded DNA with transcription factors generally contain substructures with interfaces in the same 1600 ± 400 Å^2 range of area. We called recognition modules the assembly made of these substructures and their DNA counterpart. In protein−DNA and in protein−protein recognition, the same standard-size interface is sufficient to yield a stable complex with the degree of specificity that antigen−antibody and homeodomain−DNA recognition exemplify. It is a preferred size in a number of systems.

On the DNA side, the recognition module has 12 ± 3 nucleotides, and the phosphates are a more important component than the bases. On the protein side, it has 24 ± 6 amino acid residues. Each partner in a standard-size protein−protein interface also contributes approximately that number of residues. However, the surfaces in the presence are chemically very different in the two types of complexes. Main chain atoms contribute less to protein-DNA than to protein−protein interfaces. The composition of protein−protein interfaces is somewhat less polar than the rest of the protein surface, but otherwise similar. In line with the nature of the DNA surface, the protein surface that contacts DNA is more polar than the average accessible surface. Its amino acid composition is specific and highly enriched in the basic amino acids Arg and Lys. More polar groups are buried than at a protein−protein interface of the same size, and as a consequence, the average number of hydrogen bonds per unit area is larger. In high-resolution X-ray structures, there is about one hydrogen bond per 125 Å^2 in complexes with DNA, and one per 170 Å^2 in protein−protein complexes (*22*). A standard-size interface contains 13 hydrogen bonds in one case, and about 10 in the other, to which equivalent numbers of water bridges should be added in both cases. In addition, the average protein−DNA interface is larger than the average protein−protein interface, and the total number of polar interactions is correspondingly larger. Albeit more numerous, protein−DNA hydrogen bonds are chemically less diverse than protein−protein hydrogen bonds. On the DNA side, 60% are made with the phosphate group and half of the remainder with the guanine base. On the protein side, 40% are from Lys/Arg side chains, which make more polar interactions than all other side chains together. The contribution of the peptide group is significant, but much less than at protein−protein interfaces, where it is involved in nearly two-thirds of the hydrogen bonds (*22*).

Conformation changes are an important feature of recognition. However, they are less frequent in protein−protein than in protein−DNA recognition. The components of most protease−inhibitor and antigen−antibody complexes undergo little change as can be shown by comparing their structure in the complexes and as free molecules (*39*). In contrast, the formation of complexes with large interfaces involves conformation changes, which affect the energetics of association (*40*). Interfaces with $B > 2000$ Å^2 are a minority in protein−protein complexes, but there are a sufficient number of examples to demonstrate the correlation (*22*). In the bacterial elongation factor EF-Tu−EF-Ts complex (*41*) or in transducin, a heterotrimeric G-protein (*42*), extensive changes are seen to accompany the formation of large interfaces: disordered segments of polypeptide chain become ordered, domains move, and loops rearrange on the protein surface. The same types of changes are observed in the protein component of many of the complexes with DNA analyzed here, suggesting that conformation changes correlate with the presence of large interfaces in these systems too. In addition, the amplitude of the changes that affect the DNA component also correlate to some extent with the size of the interface.

Rigid-body association involves the recognition of two surfaces with preformed complementary shapes. In contrast, the formation of an interface in elongation factors, transducin, or most protein−DNA complexes involves a large amount of induced fit. In proteins, rigid-body association can only be an approximation and local adjustments are always observed. They comprise side chain rotations on the protein surface and small deformations of the polypeptide chain. These accompany the formation of standard-size interfaces. When large interfaces are formed, global deformations are also observed, which are energetically more costly, because they perturb the packing of the protein interior. In DNA, the equivalent of a side chain rotation would be base flipping, which breaks a Watson−Crick base pair. In our sample, this is observed for the cytosine covalently bound to the active site of *Hha*I methyl transferase (*43*). In contrast, the sugar−phosphate backbone deforms easily. Moreover, some nucleotide sequences may deviate significantly from the canonical B-DNA structure we used as a reference (*1*, *44*). The well-established flexibility of DNA is apparent in the RMSD values of Table 9. The DNA backbone undergoes major changes when it binds the TBP, but the pairing of the bases in the TATA box is maintained and the protein backbone remains nearly rigid. The TBP offers the DNA a surface that is much less polar (70% nonpolar) than other proteins in our sample, and this may contribute to the distortion. In other systems, the flexibility of DNA enables the major and minor grooves to respond to the insertion of a protein component by bending in either direction (*45*−*47*).

We used the packing density of interface atoms to estimate shape complementarity in the assembled surfaces, and took the protein interior as a reference of close packing. In this way, we could show that protein−protein interfaces are close-packed whether they assemble as rigid bodies or involve conformation changes (*22*). We can now draw the same conclusion for the protein component of protein-DNA complexes, and we expect that it also applies to the DNA component, though we cannot demonstrate it at present. At protein−protein and in protein−DNA interfaces, water molecules are immobilized and can be located by crystallography. We found these water molecules to be an important element of the interfaces. Close packing largely relies on their presence. They fill cavities or, as in the TBP complex of Figure 4, they surround the region where protein and DNA atoms in contact are fully buried, a region where the highly nonpolar protein surface makes contact with the sugar moiety of the DNA. Water-mediated interactions extend this region, so that shape complementarity needs to be exact only on part of the interface.

These features of protein−DNA recognition depend neither on the exact DNA sequence nor on the type of protein that interacts with it. With few exceptions, they can be identified in the 65 complexes containing double-stranded DNA. The present sample of complexes with single-stranded DNA or RNA is too small to draw any conclusions. However, the parallel with protein−protein recognition is compelling, and we should expect rules that are common to these two processes to extend to the recognition of other biological macromolecules by proteins.

ACKNOWLEDGMENT

We thank C. Chothia for discussion, and S.E. Phillips, B. Luisi, and Y. Timsit for critical reading of the manuscript.

REFERENCES

1. Travers, A. A. (1993) *DNA-Protein Interactions*, Chapman and Hall, London, U.K.
2. Lilley, D. M. J. (1995) in *DNA-Protein: Structural Interactions* (Lilley, D. M. J., Ed.) pp 114−140, IRL Press, Oxford, U.K.
3. Freemont, P. S., Lane, A. N., and Sanderson, M. R. (1991) *Biochem. J. 278*, 1−23.
4. Suzuki, M., Brenner, S. E., Gerstein, M., and Yagi, N. (1995) *Protein Eng. 8*, 319−328.
5. Suzuki, M., and Gerstein, M. (1995) *Proteins 23*, 525−535.
6. Harrison, S. C. (1991) *Nature 353*, 715−719.
7. Pabo, C. O., and Sauer, R. T. (1992) *Annu. Rev. Biochem. 61*, 1053−1095.
8. Steitz, T. A. (1993) *Structural studies of protein-nucleic acid interaction*, Cambridge University Press, Cambridge, U.K.
9. Luisi, B. F. (1995) in *DNA-Protein: Structural Interactions* (Lilley, D. M. J., Ed.) pp 1−48, IRL Press, Oxford, U.K.
10. Bernstein, F. C., Koetzle, T. F., Williams, J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) *J. Mol. Biol. 112*, 535−542.
11. Chuprina, V. P., Rullmann, J. A., Lamerichs, R. M., van Boom, J. H., Boelens, R., and Kaptein, R. (1993) *J. Mol. Biol. 234*, 446−462.
12. Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G., and Lu, P. (1996) *Science 271*, 1247−1254.
13. Alard, P. (1992) *Calcul de surface et d'energie dans le domaine des macromolecules*, Ph.D. Thesis, Universite Libre de Bruxelles.
14. Lee, B. K., and Richards, F. M. (1971) *J. Mol. Biol. 55*, 379−400.
15. Pontius, J. (1997) *Atomic Volumes in protein crystallographic structures and their use in structure validation*, Ph.D. Thesis, Universite Libre de Bruxelles.
16. Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994) *Science 265*, 346−355.
17. Nikolov, D. B., and Burley, S. K. (1994) *Nat. Struct. Biol. 1*, 621−637.
18. Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I., and Aggarwal, A. K. (1997) *Nature 388*, 97−100.
19. Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G., and Steitz, T. A. (1985) *Nature 313*, 762−766.
20. Yang, W., and Steitz, T. A. (1995) *Cell 82*, 193−207.
21. Janin, J., Miller S., and Chothia C. (1988) *J. Mol. Biol. 204*, 155−164.
22. Lo Conte, L., Chothia, C., and Janin, J. (1999) *J. Mol. Biol. 285*, 2177−2198.
23. Mandel-Gutfreund, Y., Margalit, H., Jernigan, R. L., and Zhurkin, V. B. (1998) *J. Mol. Biol. 277*, 1129−1140.
24. Jones, S., and Thornton, J. M. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 13−20.
25. Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995) *J. Mol. Biol. 253*, 370−388.
26. McDonald, I. K., and Thornton, J. M. (1994) *J. Mol. Biol. 238*, 777−793.
27. Hunter, C. A. (1993) *J. Mol. Biol. 230*, 1025−1054.
28. Seeman, N. C., Rosemberg, J. M., and Rich, A. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 804−808.
29. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., and Sigler, P. B. (1988) *Nature 335*, 321−329.
30. Richards, F. M. (1974) *J. Mol. Biol. 82*, 1−14.
31. Harpaz, Y., Gerstein, M., and Chothia, C. (1994) *Structure 2*, 641−649.
32. Chothia, C., and Lesk, A. M. (1986) *EMBO J. 4*, 823−826.
33. Newman, M., Strzelecka, T., Dorner, L. F., Schildkraut, L., and Aggarwal, A. K. (1995) S*cience 269*, 656−663.
34. Beamer, L. J., and Pabo, C. O. (1992) *J. Mol. Biol. 227*, 177−196.
35. Kim, Y. C., Geiger, J. H., Hahn, S., and Sigler, P. B. (1993) *Nature 365*, 512−520.
36. Kim, J. L., Nikolov, D. B., and Burley, S. K. (1993) *Nature 365*, 520−527.
37. Schultz, S. C., Shields, G. C., and Steitz, T. A. (1991) *Science 253*, 1001−1007.
38. Werner, M. H., and Burley, S. K. (1997) *Cell 88*, 733−736.
39. Janin, J., and Chothia, C. (1990) *J. Biol. Chem. 265*, 16027−16030.
40. Spolar, R. S., and Record, M. T., Jr. (1994) *Science 263*, 777−784.
41. Kawashima, T., BerthetColominas, C., Wulff, M., Cusack, S., and Leberman, R. (1996) *Nature 379*, 511−518.
42. Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E., and Sigler, P. B. (1996) *Nature 379*, 311−319.
43. Reinisch, K. M, Chen, L., Verdine, G. L., and Lipscomb, W. N. (1995) *Cell 82*, 143−153.
44. Travers, A. A. (1995) in *DNA-Protein: Structural Interactions* (Lilley, D. M. J., Ed.) pp 49−75, IRL Press, Oxford, U.K.
45. Suzuki, M., Gerstein, M., and Yagi, N. (1994) *Nucleic Acids Res. 22*, 3397−3405.
46. Kyogoku, Y., Kojima, C., Lee, S. J., Tochio, H., Susuki, N., Matsuo, H., and Shirakawa, M. (1995) *Methods Enzymol. 261*, 524−541.
47. Susuki, M., and Yagi, N. (1996) *J. Mol. Biol. 255*, 677−687.
48. Sawaya, M. R., Prasad, R., Wilson, S. H., Kraut, J., and Pelletier, H. (1997) *Biochemistry 36*, 11205−11215.
49. Beese, L. S., Derbyshire, V., and Steitz, T. A. (1993) *Science 260*, 352−355.
50. Eom, S. H., Wang, J., and Steitz, T. A. (1996) *Nature 382*, 278−281.
51. Doublié, S., Tabor, S., Long, A. M., Richardson, C. C., and Ellenberger, T. (1998) *Nature 391*, 251−258.
52. Weston, S. A., Lahm, A., and Suck, D. (1992) *J. Mol. Biol. 226*, 1237−1256.
53. Wilkosz, P. A., Chandrasekhar, K., and Rosenberg, J. M. (1995) *Acta C rystallogr. Sect. D 51*, 938−945.
54. Cheng, X., Balendiran, K., Schildkraut, I., and Anderson, J. E. (1994) *EMBO J. 13*, 3927−3935.
55. Kostrewa, D., and Winkler, F. K. (1995) *Biochemistry 34*, 683−696.
56. Vassylyev, D. G., Kashiwagi, T., Mikami, Y., Ariyoshi, M., Iwai, S., Ohtsuka, E., and Morikawa, K. (1995) *Cell 83*, 773−782.
57. Feng, J. A., Johnson, R. C., and Dickerson, R. E. (1994) *Science 263*, 348−355.
58. Klimasauskas, S., Kumar, S., Roberts, R. J. and Cheng, X. (1994) *Cell 76*, 357−369.
59. Van Pouderoyen, G., Ketting, R. F., Perrakis, A., Plasterk, R. H. A., and Sixma, T. K. (1997) *EMBO J. 16*, 6044−6054.
60. Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N., and Morikawa, K. (1996) *Nature 383*, 598−603.
61. Konig, P., Giraldo, R., Chapman, L., and Rhodes, D. (1996) *Cell 85*, 125−136.
62. Rice, P. A., Yang, S. W., Mizuuchi, K., and Nash, H. A. (1996) *Cell 87*, 1295−1306.
63. Somers, W. S., and Phillips, S. E. (1992) *Nature 359*, 387−393.
64. Raumann, B. E., Rould, M. A., Pabo, C. O., and Sauer, R. T. (1994) *Nature 367*, 754−757.

65. Rodgers, D. W., and Harrison, S. C. (1993) *Structure 1*, 227−240.
66. Parkinson, G., Gunasekera, A., Vojtechovsky, J., Zhang, X. P., Kunkel, T. A., Berman, H., and Ebright, R. H. (1996) *Nat. Struct. Biol. 3*, 837−841.
67. Lawson, C. L., and Carey, J. (1993) *Nature 366*, 178−182.
68. Schumacher, M. A., Glasfeld, A., Zalkin, H., and Brennan, R. G. (1997) *J. Biol. Chem. 272*, 22648−22653.
69. Mondragon, A., and Harrison, S. C. (1991) *J. Mol. Biol. 219*, 321−334.
70. Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., and Pabo, C. O. (1991) *Cell 67*, 517−528.
71. Wilson, D. S., Guenther, B., Desplan, C., and Kuriyan, J. (1995) *Cell 82*, 709−719.
72. Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B., and Pabo, C. O. (1990) *Cell 63*, 579−590.
73. Klemm, J. D., Rould, M. A., Aurora, R., Herr, W., and Pabo, C. O. (1994) *Cell 77*, 21−32.
74. Xu, W., Rould, M. A., Jun, S., Desplan, C., and Pabo, C. O. (1995) *Cell 80*, 639−650.
75. Li, T., Stark, M. R., Johnson, A. D., and Wolberger, C. (1995) *Science 270*, 262−269.
76. Ferre d'Amare, A. R., Prendergast, G. C., Ziff, E. B., and Burley, S. K. (1993) *Nature 363*, 38−45.
77. Ferre d'Amare, A. R., Pognonec, P., Roeder, R. G., and Burley, S. K. (1994) *EMBO J. 13*, 180−189.
78. Glover, J. N., and Harrison, S. C. (1995) *Nature 373*, 257−261.
79. Brownlie, P., Ceska, T. A., Lamers, M., Romier, C., Stier, G., Teo, H., and Suck, D. (1997) *Structure 5*, 509−520.
80. Ma, P. C., Rould, M. A., Weintraub, H., and Pabo, C. O. (1994) *Cell 77*, 451−459.
81. Kodandapani, R., Pio, F., Ni, C. Z., Piccialli, G., Klemsz, M., McKercher, S., Maki, R. A., and Ely, K. R. (1996) *Nature 380*, 456−460.
82. Pellegrini, L., Tan, S., and Richmond, T. J. (1995) *Nature 376*, 490−498.
83. Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. (1992) *Cell 71*, 1223−1237.
84. Keller, W., Konig, P., and Richmond, T. J. (1995) *J. Mol. Biol. 254*, 657−667.
85. Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. (1996) *Structure 4*, 1171−1180.
86. Kim, C. A., and Berg, J. M. (1996) *Nat. Struct. Biol. 3*, 940−945.
87. Houbaviy, H. B., Usheva, A., Shenk, T., and Burley, S. K. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 13577−13582.
88. Fairall, L., Schwabe, J. W., Chapman, L., Finch, J. T., and Rhodes, D. (1993) *Nature 366*, 483−487.
89. Pavletich, N. P., and Pabo, C. O. (1993) *Science 261*, 1701−1707.
90. Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S. C. (1992) *Nature 356*, 408−414.
91. Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., and Sigler, P. B. (1991) *Nature 352*, 497−505.
92. Schwabe, J. W., Chapman, L., Finch, J. T., and Rhodes, D. (1993) *Cell 75*, 567−578.
93. Gewirth, D. T., and Sigler, P. B. (1995) *Nat. Struct. Biol. 2*, 386−394.
94. Marmorstein, R., and Harrison, S. C. (1994) *Genes Dev. 8*, 2504−2512.
95. Swaminathan, K., Flynn, P., Reece, R. J., and Marmorstein, R. (1997) *Nat. Struct. Biol. 4*, 751−759.
96. Rastinejad, F., Perlmann, T., Evans, R. M., and Sigler, P. B. (1995) *Nature 375*, 203−211.
97. Kosa, P. F., Ghosh, G., DeDecker, B. S., and Sigler, P. B. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 6042−6047.
98. Nikolov, D. B., Chen, H., Halay, E. D., Hoffman, A., Roeder, R. G., and Burley, S. K. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 4862−4867.
99. Nikolov, D. B., Chen, H., Halay, E. D., Usheva, A. A., Hisatake, K., Lee, D. K., Roeder, R. G., and Burley, S. K. (1995) *Nature 377*, 119−128.
100. Tan, S., Hunziker, Y., Sargent, D. F., and Richmond, T. J. (1996) *Nature 381*, 127−134.
101. Cramer, P., Larson, C. J., Verdine, G. L., and Muller, C. W. (1997) *EMBO J. 16*, 7078−7090.
102. Gosh, G., van Duyne, G., Ghosh, S., and Sigler, P. B. (1995) *Nature 373*, 303−310.
103. Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L., and Harrison, S. C. (1995) *Nature 373*, 311−317.
104. Muller, C. W., and Herrmann, B. G. (1997) *Nature 389*, 884−888.
105. Hegde, R. S., Grossman, S. R., Laimins, L. A., and Sigler, P. B. (1992) *Nature 359*, 505−512.
106. Buckle, A. M., and Fersht, A. R. (1994) *Biochemistry 33*, 1644−1653.
107. Padmanabhan, K., Padmanabhan, K. P., Ferrara, J. D., Sadler, J. E., and Tulinsky, A. (1993) *J. Biol. Chem. 268*, 17651−17654.
108. Bochkarev, A., Pfuetzner, R. A., Edwards, A. M., and Frappier, L. (1997) *Nature 385*, 176−181.
109. Korolev, S., Hsieh, J., Gauss, G. H., Lohman, T. M., and Waksman, G. (1997) *Cell 90*, 635−647.
110. Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J. C., and Moras, D. (1991) *Science 252*, 1682−1689.
111. Rould, M. A., Perona, J. J., and Steitz, T. A. (1991) *Nature 352*, 213−218.
112. Biou, V., Yaremchuk, A., Tukalo, M., and Cusack, S. (1994) *Science 263*, 1404−1410.
113. Nissen, P., Kjeldgaard, M., Thirup, S., Polekhina, G., Reshetnikova, L., Clark, B. F., and Nyborg, J. (1995) *Science 270*, 1464−1472.
114. Oubridge, C., Ito, N., Evans, P. R., Teo, C. H., and Nagai, K. (1994) *Nature 372*, 432−438.
115. Valegard, K., Murray, J. B., Stonehouse, N. J., van den Worm, S., Stockley, P. G., and Liljas, L. (1997) *J. Mol. Biol. 270*, 724−738.
116. Nicholls, A., Sharp, K. A., and Honig, B. (1991) *Proteins 11*, 281−296.
117. Kraulis, P. J. (1991) *J. Appl. Crystallogr. 24*, 946−950.

BI982362D