3362–3376 *Nucleic Acids Research, 2001, Vol. 29, No. 16*

Standard atomic volumes in double-stranded DNA and packing in protein–DNA interfaces

Katalin Nadassy[1,2], Isabel Tomás-Oliveira[3], Ian Alberts[1,2], Joël Janin[1,4] and Shoshana J. Wodak[1,3,*]

[1]European Bioinformatics Institute, EMBL, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Department of Biological Sciences, University of Stirling, Stirling FK9 4LA, UK, [3]SCMB, Université Libre de Bruxelles, CP 160/16, Avenue F. D. Roosevelt, B-1050-Bruxelles, Belgium and [4]Laboratoire d'Enzymologie et de Biochimie Structurales, CNRS UPR9063, 91198-Gif-sur-Yvette, France

Received April 30, 2001; Revised and Accepted July 6, 2001

ABSTRACT

Standard volumes for atoms in double-stranded B-DNA are derived using high resolution crystal structures from the Nucleic Acid Database (NDB) and compared with corresponding values derived from crystal structures of small organic compounds in the Cambridge Structural Database (CSD). Two different methods are used to compute these volumes: the classical Voronoi method, which does not depend on the size of atoms, and the related Radical Planes method which does. Results show that atomic groups buried in the interior of double-stranded DNA are, on average, more tightly packed than in related small molecules in the CSD. The packing efficiency of DNA atoms at the interfaces of 25 high resolution protein–DNA complexes is determined by computing the ratios between the volumes of interfacial DNA atoms and the corresponding standard volumes. These ratios are found to be close to unity, indicating that the DNA atoms at protein–DNA interfaces are as closely packed as in crystals of B-DNA. Analogous volume ratios, computed for buried protein atoms, are also near unity, confirming our earlier conclusions that the packing efficiency of these atoms is similar to that in the protein interior. In addition, we examine the number, volume and solvent occupation of cavities located at the protein–DNA interfaces and compared them with those in the protein interior. Cavities are found to be ubiquitous in the interfaces as well as inside the protein moieties. The frequency of solvent occupation of cavities is however higher in the interfaces, indicating that those are more hydrated than protein interiors. Lastly, we compare our results with those obtained using two different measures of shape complementarity of the analysed interfaces, and find that the correlation between our volume ratios and these measures, as well as between the measures themselves, is weak. Our results indicate that a tightly packed environment made up of DNA, protein and solvent atoms plays a significant role in protein–DNA recognition.

INTRODUCTION

The volumes of atoms and residues and their packing inside proteins have been of interest because they have an important bearing on the physical and thermodynamic properties of the folded polypeptide. Optimal side chain packing inside proteins is believed to be key in stabilizing the native state of proteins (1,2). Achieving or disrupting it is thought to govern the barriers of the folding and unfolding processes, respectively (3,4). Substitutions of a larger residue by a smaller one, which create empty space inside the protein, are destabilizing (5).

The availability of protein crystal structures has made it possible to evaluate the volumes and packing densities of atoms and residues in proteins from the atomic coordinates. An accurate method for computing these quantities in molecular systems involves partitioning space between atoms by building the so-called Voronoi polyhedra. In the classical method of Voronoi (6), the faces of the polyhedra, or the dividing planes, are positioned exactly halfway between neighboring atoms. In the related Radical Planes method (7) these planes are positioned in a manner proportional to the atomic radii, whereas in the Richards-B method (8,9), the dividing planes are positioned differently between bonded atoms than between non-bonded ones. The latter two variants require the use of atomic radii. The early work (8) used the radii derived by Bondi (10). Later work was based on new radii, derived by Chothia and Janin (11) and more recently by Li and Nussinov (12) and Tsai *et al.* (13).

The different variants of the Voronoi method have been used to compute standard volumes for atoms and residues in proteins, first in a limited set of structures (8,11,14), and later on much larger sets (13,15,16). An important finding of these studies was that in general, residues in the interior of proteins

*To whom correspondence should be addressed at: SCMB, Université Libre de Bruxelles, CP 160/16, Avenue F. D. Roosevelt, B-1050-Bruxelles, Belgium. Tel: +32 2 648 5200; +32 2 648 8954; Email: shosh@ucmb.ulb.ac.be
Present address:
Katalin Nadassy, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

are close packed, with mean volumes equal to (8), or somewhat smaller than those they have in amino acid crystals (13,15). They also showed that packing differs in different regions of the protein (8,16–19). In particular, it was found that the buried protein core is on average more tightly packed than regions on the surface (19,20).

With the significant increase in the available data on structures of macromolecular complexes and the mounting interest in improving our understanding of molecular recognition, volume calculations of the type described above have been applied to analyze the packing of atoms in protein–protein and protein–DNA interfaces. Conte *et al.* (21), computed the volumes of atoms buried in interfaces of protein–protein complexes and showed that they were similar to those inside proteins, therefore, concluding that these interfaces were, on average, as closely packed as the protein interior.

A similar analysis performed by some of us for protein atoms buried in protein–nucleic acid complexes (22), led to analogous conclusions about the protein atoms in these interfaces. Moreover, that analysis showed that performing the volume calculations in the presence of the crystallographic water molecules, yielded a better agreement between the volumes of interfacial atoms and those in the protein interior, suggesting that water molecules play an important role in fostering optimal packing in these interfaces. But the packing efficiency of nucleic acid atoms was not evaluated in that study due to the lack of a standard set of atomic volumes and radii for nucleic acid atoms.

In this work, we extend the analysis of atomic volumes and packing to nucleic acid atoms, more particularly to DNA. We derive the mean volumes and standard deviations for atoms in DNA using crystal structures of various DNA forms in the Nucleic acid Database (NDB) (23). These volume distributions are compared to those derived for nucleic acid groups in crystals of small molecules from the Cambridge Structural Database (CSD) (24). Using the standard volumes computed for atoms in B-DNA as the reference, we evaluate the packing of DNA atoms at the interfaces of 25 protein–DNA complexes determined at high resolution (better than 2.4 Å). In addition to atomic volume calculations, we also compute the volume of empty and water-filled cavities in these interfaces.

Results show that DNA atoms in double helical DNA structures are on average more closely packed than the equivalent atoms in small molecule crystals, an analogous observation to that made previously for amino acids (15). We show furthermore that DNA atoms in protein–DNA interfaces are, on average, as closely packed as in crystals of B-DNA, and here too, water molecules play a crucial role in fostering close packing. This picture is largely confirmed by our analysis of empty and filled cavities.

Two other measures for assessing packing in interfaces of macromolecular complexes have been proposed. One is the shape correlation index of Lawrence and Colman (25) and the other is the gap volume of Jones and Thornton (26), used recently to evaluate complementarity in protein–DNA interfaces (27). We apply both measures to our dataset of 25 protein–DNA complexes, and the results are compared with those obtained with the volume and cavity calculations.

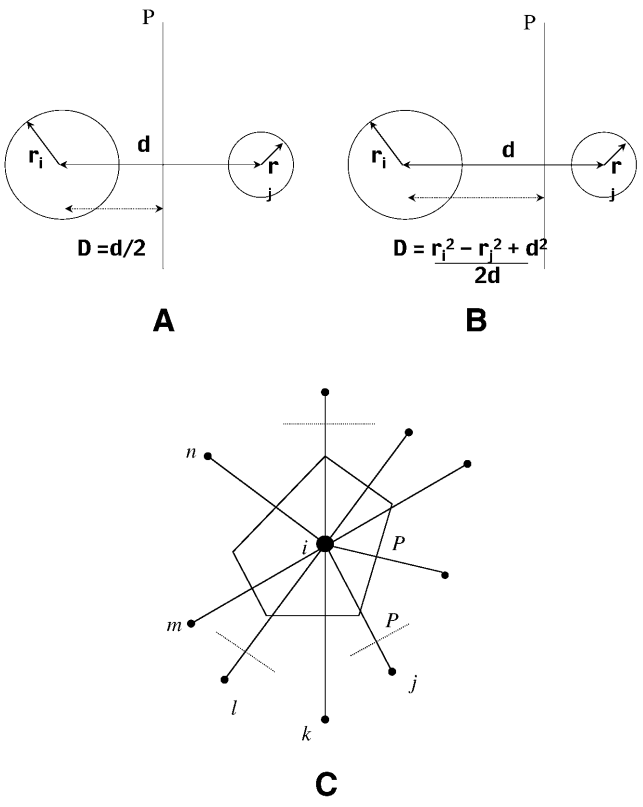

Figure 1. Illustration of the Voronoi procedure for computing atomic volumes. (**A**) The classical Voronoi procedure, where the space between atoms i and j, separated by a distance d, is partitioned by positioning the dividing plane P at a distance D = d/2, exactly midway between the two atoms. (**B**) The Radical Planes method, where plane P is positioned at a distance $D = (r_i^2 - r_j^2 + d^2)/2d$ from atom i. This plane represents the geometric locus of the points of intersection of the equidistant tangents drawn to the van der Waals spheres of the atoms (7). (**C**) 2D representation of the Voronoi polyhedron around a central atom. Lines are constructed from the central atom (i) to its neighbors (j, k, l …) and the planes P are positioned perpendicular to the vectors according to the Voronoi or Radical Planes methods as in (A) or (B). The volume of the atom is defined as the volume of the smallest polyhedron delimited by the planes P.

MATERIALS AND METHODS

Volume calculations

The volume calculations for both the Voronoi and Radical Planes methods were conducted using the SURVOL (28) software package. In these methods, Voronoi polyhedra are constructed around each atom with the dividing planes positioned either halfway between neighboring atoms (Voronoi) or in proportion to the VDW radii of neighboring atoms (Radical Planes). The volume assigned to each atom is given by the space occupied by the smallest polyhedron defined by these planes (see Fig. 1). The Voronoi polyhedra are defined only for atoms that are completely surrounded by other atoms. To identify such buried atoms, the solvent accessible surface area was computed with a probe size of 1.5 Å, using the program SURVOL (28), which implements a modified version of the method of Connolly (29). In cases where the PDB structure contained more than one protein–DNA interface in the

Table 1. Dataset of high resolution protein–DNA complexes

PDB Code	complex	Resolution (Å)	Chain code and number of Amino acids	Chain code and number of nucleotides
1bpy	DNA polymerase	2.2	A335	T16, P10, D5
1t7p	Phage T7	2.2	A698, B108	P11, T13
1bhm	BamHI	2.2	A213, B213	C12, D12
1dnk	DNase I	2.3	A260	B7, C8
1rvc	EcoRV	2.1	A244, B244	C6, D5, E6, F5
1hcr	Hin recombinase	2.3	A52	B13, C14
1tc3	C. elegans transposase	2.4	C51	A21, B20
1ign	RAP1 telomere binding	2.3	A246	C19, D19
1lmb	Lambda repressor	1.8	3:92, 4:92	1:20, 2:20
1tro	Trp repressor	1.9	A108, C108	I19, J19
1trr	Trp repressor, half-site	2.4	A107, B107	C16, I16
1fjl	Paired dimer	2	A81, B81	D14, E14
1pue	Pu1-ETS domain	2.1	E89	A16, B16
2dgc	GCN4, ATF site	2.2	A63	B19
1aay	Zif 268	1.6	A90	B11, C11
1mey	Designed	2.2	C87	A13, B13
1hcq	Estrogen receptor	2.4	A84, B84	C18, D18
1lat	Glucocorticoid	1.9	A82	C19, D19
2nll	Retinoid receptor	1.9	A66, B103	C18, D18
1ais	TBP-TFIIB	2.1	A182, B200	C17, E17
1cdw	TBP,human	1.9	A179	B16, C16
1a3q	NF_-B p52	2.1	A285, B285	C11, D11
1nfk	NF_-B p50	2.3	A325, B325	C11, D11
1tsr	p53 core	2.2	B219	E21, F21
2bop	E2 domain	1.7	A85	B17

This set of protein–DNA complexes is the same as that used in Nadassy *et al.* (22). The two structures from the trp repressor proteins were included, because one (1trr) is the half-site tandem complex, and the two structures differ in resolution, the number of residues adopting a well defined conformation in the crystal, and the length of the DNA chains. It seemed worthwhile to check how these differences influence the results.

asymmetric unit only one copy of this interface was used. The chain name and residue numbers for each protein–DNA complex are listed in Table 1.

Radii of atomic groups

The Radical Planes method requires the assignment of radii to atoms or atomic groups. In this work the united atom approximation is used, in which heavy atoms are considered together with their bound hydrogens. Several sets of united atom groups have been defined for volume calculations in proteins. We use the definition of Tsai *et al*. (13) and include three additional atomic groups for nucleic acids that are not found in proteins as listed in Table 2. Tsai *et al*. determined radii for the atomic groups in proteins from a detailed analysis of the radial atom pair distributions as a function of distance in small molecule crystal structures in the CSD (24). We used the same procedure to derive values for the three additional groups in nucleic acids. The list of structures used for these calculations was the same as that used in deriving the standard volumes (see below) and can be found in the Supplementary Material. For Zn, I and Mg, found in the CSD, we used the radii of Bondi (10).

Cavity calculations

A cavity is defined here as a region of empty space completely surrounded by protein or DNA atoms, whose volume is delimited by the so-called molecular surface (9). Cavity locations and volumes were computed using the software SURVOL (28), which implements the algorithm by Alard and Wodak (30). Two calculations were carried out. One in the absence of crystallographic water molecules and a second in their presence. Cavities which are identified in the first calculation but not in the second, are classified as water-filled, whereas the cavities identified in the presence of water molecules are classified as empty. A probe radius of 1.4 Å was used.

The PDB entry 1bmh, which has missing side chains for 22 of its residues, was excluded from the cavity calculation. In other structures with some missing side-chain atoms, cavities lined by these side chains were excluded from the analysis.

Gap index and shape complementarity calculations

To complement our volume calculations we computed two additional quantities previously proposed for evaluating the extent of packing in inter-molecular interfaces. One quantity is the 'gap volume index' of Laskowski (31), used by Jones and Thornton (26) and Jones *et al.* (27). This index is defined as the available volume or 'gap volume' between the solvent accessible surfaces of the molecular components of the complex, divided by the surface area buried in the interface. The gap volume is calculated as follows (see http://www.biochem.ucl.ac.uk/bsm/DNA/server/parameter_def.html): for all pairs of atoms, belonging respectively to the protein and DNA molecules, the size of the largest sphere (maximum diameter 10 Å) that can be placed midway between the surfaces of the two atoms, without penetrating the van der

Table 2. Standard radii for atomic groups in DNA[a]

Atom type[b]	*VDW radii*[c]
C3H0	1.61
C3H1	1.76
C4H1	1.88
C4H2	1.88
C4H3	1.88
N2H0[d]	1.61
N3H0	1.64
N3H1	1.64
N3H2	1.64
N4H3	1.64
O1H0	1.42
O2H1	1.46
O2H0[d]	1.51
P4H0[d]	2.04

[a]We used the radii derived by Tsai *et al*. (13) except for three atomic groups.
[b]Following Tsai *et al*. (13) atomic groups are given the label *AnHm*, where *A* is the non-hydrogen atom, *n* its valence and *m* the number of directly bonded hydrogen atoms, e.g. C_4H_2 refers to a tetrahedral carbon atom covalently bonded to two hydrogen atoms and two non-hydrogen atoms.
[c]Radii in Å.
[d]Radii for these groups were determined in our group.
The list of structures from the CSD, used for computing the radii, can be found in the Supplementary Material.

Waals surface of any other atom (the gap-sphere), is determined. Spheres whose radius is <1Å are discarded. The gap volume is obtained by summing the volumes of all allowable gap-spheres. Using this algorithm, the largest gap spheres will be those in contact with the surface of atoms belonging to regions on the periphery of the interface that are accessible to the bulk solvent in the complex. The number of these peripheral spheres would essentially depend on the size of the interface, and they should therefore contribute more to the gap volume than other gap spheres placed in the middle of the interface.

The second evaluated quantity is the shape correlation index (S_c) of Lawrence and Colman (25), derived from the distance between points on the surfaces of the interacting molecules and the angle between the vectors normal to these surfaces.

The gap volume index for our sample of protein–DNA interfaces was evaluated using software provided by the Web server of Thornton and colleagues (http://www.biochem.ucl.ac.uk/bsm/DNA/server/). The shape correlation index S_c was computed using software kindly provided by the authors.

Structural datasets

Nucleic acid structures from the NDB. An appropriate subset of high-resolution double-stranded B-DNA structures was selected from the NDB on the basis of the following criteria. The resolution had to be 2.0 Å or better, and the R-factor ≤20%. In addition, the phosphate atoms of the structure were required to display a root mean square deviation (r.m.s.d.) of ≤1.60 Å upon superposition onto a 40 base pair canonical B-DNA structure (A.R.Srinivasan and W.K.Olson, Rutgers State University of New Jersey, personal communication). This yielded the 35 double-stranded B-DNA structures, analyzed in this study. A complete list of these structures can be found in the Supplementary Material.

Datasets for the A and Z conformations of DNA were selected using the same criteria, with the exception that superposition was not carried out on these structures. This produced 46 A-DNA and 34 Z-DNA structures. For a complete list of these structures see the Supplementary Material.

Small molecule structures from the Cambridge Structural Database (CSD). To select nucleic acid structures from the CSD, we used the program QUEST (32). Independent searches were performed for CSD entries that contained nucleic acid groups corresponding to the four different bases, sugar and phosphate groups, respectively. For each search we considered entries with R < 6% and which were devoid of valence errors, large bond length deviations or disorder. The codes of the retrieved CSD structures are listed in the Supplementary Material.

In order to conduct the volume calculations on a sufficient number of buried atoms, neighboring molecules in the crystal were generated using the program FILLXR (24). The retrieved CSD structures generally contain atoms and atomic groups that are not found in either proteins or nucleic acids. Since consistent radii have not been assigned to these atomic groups in the manner of Tsai *et al*. (13), only the classical Voronoi procedure was used to compute atomic volumes in the CSD structures.

Structures of protein–DNA complexes. The dataset of protein–DNA complexes was the same as that used in our previous study (22). It consists of 25 complexes of proteins with double-stranded DNA, from the protein databank (PDB), with resolution of 2.4 Å or better, listed in Table 1. The analysis was performed on the biologically relevant assemblies (monomers or dimers) as described previously (22) and also listed in Table 1.

RESULTS

Atomic volumes in B-DNA structures from the NDB

Table 3 lists the mean volumes and standard deviation of the atomic groups in the four bases and the sugar–phosphate group, computed for the reference set of 35 B-DNA structures from the NDB. B-DNA is considered to be the canonical form of DNA. Its double helix has 10 bp/pitch, an axial rise of 3.3–3.4 Å, and contains a wide major groove and narrow minor groove. Stacking in B-DNA is limited to intra-strand interactions and the base pair planes are perpendicular to the double helix axis.

Two methods were used to compute the volumes. The classical Voronoi method, in which space is assigned by positioning the dividing panes midway between the atoms, and the Radical Planes method, which positions the dividing planes in a manner proportional to the atomic radii. The radii for nucleic acid atoms, used in the Radical Planes method, were those derived by Tsai *et al*. (13) for equivalent atom types in proteins. Radii for three additional atomic groups present in nucleic acids, but not in proteins, were computed here using the same procedure. Volumes were computed only for buried atoms, and in order to increase the number of examples the calculations were performed also in the presence of crystallographic water

Table 3. Standard reference atomic volumes for double-stranded B-DNA

	NDB Dataset Radical Planes Method			NDB Dataset Voronoi method			CSD Dataset Voronoi Method		
	Average	*St.dev.*	*Freq.*	*Average*	*St.dev.*	*Freq.*	*Average*	*St.dev.*	*Freq.*
Adenine									
C2	21.8	1.5	69	16.0	1.4	61	17.7	1.4	38
C4	8.3	0.4	128	8.7	0.4	121	9.2	0.7	38
C5	8.7	0.2	128	8.9	0.2	128	9.6	0.6	38
C6	8.0	0.3	128	8.3	0.2	122	9.2	0.7	38
C8	20.4	1.5	68	16.8	1.6	74	17.4	1.4	38
N1	11.8	0.7	124	13.7	0.9	127	15.3	1.4	38
N3	13.1	1.5	73	15.0	1.9	76	16.6	1.5	38
N6	23.8	1.7	64	21.8	1.9	67	22.9	1.7	38
N7	14.1	1.8	70	15.6	1.9	67	17.3	1.4	38
N9	6.1	0.2	129	9.0	0.5	134	9.3	0.6	38
Guanine									
C2	7.7	0.3	149	8.2	0.2	68	8.7	0.7	11
C4	8.6	0.4	148	8.9	0.4	141	9.5	0.7	11
C5	9.0	0.5	148	9.1	0.5	148	9.7	0.9	11
C6	9.9	0.4	148	8.2	0.4	133	8.9	0.7	11
C8	20.8	1.5	54	17.0	1.7	66	17.3	1.3	11
N1	14.6	0.6	142	13.8	1.1	148	14.5	0.9	11
N2	22.9	1.7	31	21.4	2.0	36	22.9	1.7	11
N3	14.7	1.2	66	14.7	1.3	65	15.8	1.1	11
N7	15.0	2.0	68	16.1	2.0	52	15.4	1.1	11
N9	6.2	0.3	155	10.3	2.4	217	9.4	0.6	11
O6	14.4	1.4	80	19.2	2.1	90	21.6	2.7	11
Cytosine									
C2	9.9	0.4	139	8.3	0.3	131	8.4	0.5	27
C4	7.0	0.3	147	8.5	0.3	131	9.0	0.9	27
C5	20.9	1.4	36	17.3	1.3	33	17.9	2.4	27
C6	18.4	1.3	79	16.1	1.3	73	16.0	1.4	27
N1	6.3	0.2	150	9.8	1.7	230	9.0	0.5	27
N3	13.7	0.8	138	13.8	1.2	144	14.8	0.8	27
N4	23.1	1.4	64	21.3	1.6	69	21.7	1.4	27
O2	14.4	1.1	48	20.3	2.4	54	20.5	1.4	27
Thymine									
C2	9.3	0.4	119	8.2	0.3	108	8.6	0.5	52
C4	9.7	0.3	118	8.3	0.3	109	9.0	0.7	52
C5	5.9	0.5	119	9.9	0.3	113	9.9	0.7	52
C5M	36.7	0.9	12	27.3	1.0	12	28.7	2.3	52
C6	19.2	1.2	77	15.9	1.2	71	15.9	1.2	52
N1	6.1	0.3	118	8.9	0.7	136	9.1	0.5	52
N3	14.8	0.8	113	13.7	1.0	121	14.7	1.1	52
O2	15.5	1.7	66	21.0	2.0	68	22.6	2.0	52
O4	15.4	1.7	86	20.5	2.2	90	22.1	1.9	52
Sugar-P									
C1*	18.1	0.9	309	11.2	0.8	282	11.1	0.6	31
C2*	21.5	1.4	184	18.0	1.4	209	19.3	1.7	31
C3*	16.1	1.3	115	11.2	1.1	60	11.8	0.7	31
C4*	16.9	1.3	62	11.5	0.9	52	11.5	0.8	31
C5*	26.2	1.7	22	19.0	1.7	26	19.9	1.8	31
O3*	7.1	1.5	175	16.3	1.9	200	20.3	2.4	31
O4*	8.8	1.3	277	17.5	3.2	357	15.8	1.2	31
O5*	7.7	1.4	295	16.4	1.8	208	18.8	3.7	31
O1P	11.6	1.1	17	20.7	0.7	11	19.7	1.2	59
O2P	11.5	1.6	20	19.9	1.7	14	20.7	2.2	51
P	29.3	0.8	10	6.4	0.2	115	6.4	0.1	176
Bases[a]									
A	136.2			133.9			144.5		
G	143.8			146.6			153.7		
C	113.6			115.6			117.2		
T	132.4			133.6			140.5		
AT	268.6			267.5			285.0		
GC	257.4			262.2			270.9		
Sugar + P	174.8			168.1			175.3		

[a]The volumes for the bases (A, G, C and T) and sugar + phosphate groups are computed by summing the volumes of the constituent atoms.

positions. This notwithstanding, the number of examples was below 20 for the OP1 and OP2 phosphate oxygens (Table 3), which are generally the most solvent exposed groups in nucleic acids.

We see that the Voronoi volumes of polar atoms tend to display larger standard deviations (13–23%) than those of non-polar ones (2–10%), presumably because they tend to be more solvent exposed, and hence surrounded by fewer non-bonded neighbors, in agreement with previous observations made in proteins (13,16,18). Variations in atomic volumes are also caused by differences in their bonding environment and in their chemical type (16,18). For example, the C1′ and C2′ sugar sp^3 carbon groups have the same atomic radius, but the former, which is bonded to three heavy atoms, has a smaller volume than the latter, which bonds to only two. Similar trends are observed for several atoms in the bases.

The Voronoi and Radical Planes volumes (Table 3) display very similar standard deviations, but the mean values computed by the two methods can differ significantly when bonded atoms have different van der Waals radii. A striking example is the phosphorous atom, which appears much larger with the Radical Planes method than with Voronoi, whereas the phosphate oxygens are smaller.

Mean and standard deviation of the volumes of the four DNA bases and sugar–phosphate groups (bottom of Table 3) were computed by summing the volumes of the constituent atoms. Remarkably, the respective volumes of the A-T and C-G base pairs differ by <4%.

Atomic volumes in A- and Z-DNA

Atomic volume calculations were also conducted for A- and Z-DNA, in order to determine the influence of the DNA conformation on the volumes of its atoms. All calculations were performed including crystal neighbors and water molecules. These DNA conformers have several distinctive features. A-DNA has 11 bp/pitch, an axial rise of ~2.6 Å, a shallower major groove, and a deeper minor groove than B-DNA. Z-DNA is a left-handed helix with 12 bp/turn and an axial rise of 3.7 Å.

Table 4 lists the mean atomic volumes and their standard deviations, computed using the Radical Planes method for our dataset of 46 A-DNA structures and 34 Z-DNA structures listed in Tables S2 and S3 in the Supplementary Material, respectively. On the whole, the listed values resemble those found in B-DNA. The atoms displaying statistically significant differences in their volume distributions in the three DNA forms are nearly all in the sugar–phosphate moiety. The displayed trends vary with the atom type. For example the volumes of the O1P, O2P and O3′ atoms in A-DNA, and those of the C2′ and O3′ atoms in Z-DNA, are larger than those of their counterparts in B-DNA. But those of the Z-DNA sugar atoms O4′ and C4′are smaller than in B-DNA (Table 4).

As for the total volumes of individual bases and the sugar–phosphate moiety, those are in general larger in A-DNA than B-DNA. The smaller volumes of Z-DNA groups is probably not significant, owing to the small number of observations, caused by the fact that these groups tend to be more solvent exposed than in the two other DNA forms.

On the basis of these considerations, we take the atomic volumes distributions for B-DNA as the standard against which DNA volumes in individual structures are compared.

The volumes of atoms in nucleic acid groups from the CSD

The CSD was searched for structures containing deoxynucleotide components, namely, the individual bases, sugar and phosphate moieties. The CSD structures identified for each type of component are listed in Table S4 in the Supplementary Material. To obtain a sufficient number of buried atoms for which the volumes can be computed, the neighboring molecules in the crystal unit cell were generated. Volume calculations were done using only the Voronoi method, in order to avoid the problem of deriving radii for a large variety of atoms in chemically diverse organic and inorganic molecules with which the nucleic acid moieties are associated in the CSD.

The resulting mean atomic volumes and standard deviations are listed in the right-most column of Table 3. The standard deviations are in the range of 2–19%. Those of the rarer buried phosphate and the O3′ and O5′ sugar oxygens are the largest (11–19%). These values are similar to the standard deviations of the Voronoi volumes of atoms in B-DNA, which range from 2 to 23%. Interestingly, with a few exceptions, the mean atomic volumes in the CSD structures are larger than those computed in double-helical DNA structures from the NDB.

Summing up the atoms of bases and sugar–phosphate groups yields volumes for these groups which are 5% larger, on average, (and 8% larger for adenine) in the CSD than in the NDB. These findings parallel the reported observations that the volumes of amino acid residues in protein cores are ~5% smaller than those in amino acids crystal structures (15). They can be readily rationalized by the fact that in the NDB, the nucleic acid moieties are part of tightly packed B-DNA structures, where base pairs are stacked and form Watson–Crick type H-bonds. In the CSD on the other hand, the same groups, particularly the charged phosphates, which display the largest volume differences, are in widely different surroundings often bearing little resemblance with those in DNA crystals. This, together with the constraints imposed by the requirement of forming a 3D lattice, leads to looser overall packing.

Packing in protein–DNA interfaces

Volumes of atoms buried in protein–DNA interfaces. In our previous study of protein–nucleic acid complexes (22), the packing efficiency of protein atoms buried at the interfaces in a set of 25 high-resolution protein–nucleic acid complexes was examined. To this end, the ratio $[V/V_0]^{PROT}$, was computed, where V is the sum of the volumes of protein atoms buried at each protein–DNA interface and V_0 is the sum of standard volumes for the same atoms. The standard volumes were taken to be those computed for atoms buried in the protein core (18). Values of $[V/V_0]^{PROT} < 1.0$ were taken to indicate that the atoms buried at the interface are more tightly packed than in the protein interior, whereas looser, or similar, packing as in the protein interior was indicated by values of $[V/V_0]^{PROT} \geq 1.0$. In that study, however, the packing of DNA atoms was not evaluated, since we did not have at our disposal a set of standard volumes and atomic radii for atoms in DNA.

With the set of standard atomic volumes in B-DNA and the atomic radii derived here, this limitation no longer exists and we can extend our packing analysis to include the evaluation of the volumes of the DNA component of the interfaces. This analysis was performed on 25 high resolution protein–DNA complexes (2.4 Å or better) from the earlier dataset of Nadassy

Table 4. Atomic volumes for A-, Z- and B-DNA[a]

	A-DNA			Z-DNA			B-DNA		
	average	stdev	freq	average	stdev	freq	average	stdev	freq
Adenine									
C2	22.7	1.4	22	23.2	1.5	2	21.8	1.5	69
C4	8.6	0.4	52	7.5	0.1	4	8.3	0.4	128
C5	9.0	0.4	52	8.7	0.3	4	8.7	0.2	128
C6	8.0	0.2	52	8.1	0.2	4	8.0	0.3	128
C8	21.1	1.0	29				20.4	1.5	68
N1	11.7	0.5	50	12.2	0.3	3	11.8	0.7	124
N3	13.4	0.9	31	12.7	0.5	4	13.1	1.5	73
N6	23.3	1.5	23	14.3	0.0	1	23.8	1.7	64
N7	14.0	1.3	32	5.9	0.2	4	14.1	1.8	70
N9	6.4	0.3	48				6.1	0.2	129
Guanine									
C2	7.6	0.3	226	7.4	0.4	85	7.7	0.3	149
C4	8.3	0.4	228	7.9	0.3	85	8.6	0.4	148
C5	9.0	0.4	226	9.1	0.3	83	9.0	0.5	148
C6	10.0	0.4	222	10.3	0.4	85	9.9	0.4	148
C8	22.4	1.4	155	23.9	2.3	2	20.8	1.5	54
N1	14.6	0.6	204	14.8	0.3	72	14.6	0.6	142
N2	23.8	1.9	24	23.7	1.3	54	22.9	1.7	31
N3	14.1	1.2	90	14.0	0.6	83	14.7	1.2	66
N7	14.5	1.7	144	16.0	1.6	13	15.0	2.0	68
N9	6.2	0.3	240	6.2	0.3	86	6.2	0.3	155
O6	14.8	1.6	143	16.2	2.1	23	14.4	1.4	80
Cystosine									
C2	9.5	0.4	224	9.8	0.3	75	9.9	0.4	139
C4	7.0	0.7	215	6.8	0.3	76	7.0	0.3	147
C5	21.4	1.7	124	20.5	1.3	31	20.9	1.4	36
C6	18.8	1.1	173	19.5	0.8	50	18.4	1.3	79
N1	6.3	0.3	230	6.3	0.2	78	6.3	0.2	150
N3	13.5	0.6	212	14.0	0.4	73	13.7	0.8	138
N4	22.9	1.6	112	23.4	1.5	23	23.1	1.4	64
O2	14.4	1.4	77	15.0	1.3	57	14.4	1.1	48
Thymine									
C2	9.2	0.4	54	9.3	0.1	10	9.3	0.4	119
C4	9.8	0.5	52	9.6	0.3	10	9.7	0.3	118
C5	6.0	0.4	52	5.5	0.2	10	5.9	0.5	119
C5M	37.9	2.0	14				36.7	0.9	12
C6	19.4	0.8	45	20.4	0.3	3	19.2	1.2	77
N1	6.2	0.3	54	6.4	0.2	10	6.1	0.3	118
N3	14.6	0.5	50	14.9	0.7	6	14.8	0.8	113
O2	16.3	0.6	14	16.5	0.0	1	15.5	1.7	66
O4	13.5	0.9	32				15.4	1.7	86
Sugar-P									
C1*	19.2	1.4	53	18.6	0.7	99	18.1	0.9	309
C2*	21.9	1.8	45	23.6	1.0	63	21.5	1.4	184
C3*	15.3	0.9	544	15.4	0.8	142	16.1	1.3	115
C4*	17.6	0.9	16	14.8	0.7	142	16.9	1.3	62
C5*	27.1	1.9	27	24.8	1.2	19	26.2	1.7	22
O3*	8.4	1.5	69	9.1	1.2	77	7.1	1.5	175
O4*	8.3	1.1	157	6.6	0.8	162	8.8	1.3	277
O5*	7.4	1.3	462	7.9	1.3	65	7.7	1.4	295
O1P	13.3	2.2	27	14.1	2.6	6	11.6	1.1	17
O2P	12.9	2.4	58	11.4	2.3	11	11.5	1.6	20
P	30.2	1.0	33	20.2	10.6	6	29.3	0.8	10
Bases[b]									
A	138.3			92.4			136.1		
G	145.4			149.4			143.8		
C	113.9			115.2			113.2		
T	132.9			82.7			132.6		
AT	271.2			175.1			268.7		
GC	259.3			264.7			257.0		
Sugar + P	181.8			166.6			174.8		

[a]Calculated using the Radical Planes method.
[b]The volumes for the bases (A, G, C and T) and sugar + phosphate groups are computed by summing the volumes of the constituent atoms.

Table 5. Packing efficiency at protein–DNA interfaces[a]

PDB Code	Complex	Number of DNA interface atoms	Absence of water		Presence of water		
			$[V/V_0]^{DNA}$	% buried	$[V/V_0]^{DNA}$	%buried	$[V/V_0]^{PROT}$
1bpy	DNA polymerase	164	1.03	30	0.99	53	1.00
1t7p	Phage T7	202	1.08	23	1.04	37	1.04
1bhm	BamHI	222	1.06	36	1.00	64	1.02
1dnk	Dnase I	84	1.06	26	1.03	40	1.02
1rvc	EcoRV	247	1.09	24	1.05	68	1.03
1hcr	Hin recombinase	190	1.02	46	0.97	47	1.02
1tc3	Transposase	118	1.00	39	0.97	49	1.01
1ign	RAP1 telomere	253	1.04	40	0.98	57	0.97
1lmb	Lambda repressor	172	1.14	26	1.04	52	1.02
1tro	Trp repressor	176	1.08	28	1.03	64	1.04
1trr	Trp repressor,half site	167	1.00	40	0.98	56	1.01
1fjl	Paired dimer	210	1.02	30	0.97	64	0.99
1pue	Pu1-ETS domain	117	1.06	20	1.01	50	1.00
2dgc	GCN4,ATF site	150	1.07	7	1.00	11	1.02
1aay	Zif268	155	1.03	22	1.01	58	1.01
1mey	Designed	145	0.99	21	1.00	56	1.01
1hcq	Estrogen receptor	130	1.04	36	0.99	55	1.04
1lat	Glucocortiocoid	79	1.06	30	0.99	75	1.00
2nll	Retinoid receptor	182	1.09	24	1.03	53	1.02
1ais	TBP-TFIIB	203	1.12	37	1.06	58	1.03
1cdw	TBP, human	175	1.12	54	1.05	75	1.03
1a3q	NFk-B p52	177	1.02	24	1.05	62	1.03
1nfk	NFk-B p50	190	1.03	23	1.06	56	1.03
1tsr	p53 core	77	1.08	34	1.04	45	1.08
2bop	E2 domain	182	1.12	11	1.02	24	1.01

[a]V is the sum of the volumes of nucleic acid atoms buried at interfaces with proteins, V_0 is the sum of standard reference volumes for atoms buried inside B-DNA. All volumes are computed using the Radical Planes method. The total number of buried nucleic acid atoms at the interface is listed and the percentage that are buried. The last column gives the re-computed protein volume ratios $[V/V_0]^{PROT}$ using the newly defined radii set of Tsai *et al.* (13).

et al. (22) (Table 1). Lower resolution structures from that dataset were not considered, in order to minimize the bias on volume values resulting from crystal structure imprecision (16,21).

In these 25 complexes, on average only 30% of the DNA atoms in the interface are completely buried in the absence of crystallographic water molecules. But this fraction rises to 54% when these molecules are included. A similar trend was observed when protein atoms of the same interfaces were analyzed (22) and in protein–protein complexes (21). Analyses of the interface volumes here and in previous studies were therefore performed both in the presence and absence of the crystallographic water molecules. All atomic volumes were computed using the Radical Planes method to preserve consistency with the values computed previously for the protein portion of the same interface (22).

Table 5 lists the volume ratios denoted $[V/V_0]^{DNA}$ computed for the DNA portions of the interfaces in our dataset of 25 complexes. Also listed are the number of atoms and the percent of buried atoms in each interface. Figure 2A displays the histograms of the $[V/V_0]^{DNA}$ values.

In the absence of the interfacial water molecules, the mean $[V/V_0]^{DNA}$ ratio for the 25 complexes is 1.06, indicating that the DNA atoms in the interface are on average less well packed than in B-DNA. Figure 2A shows that the individual $[V/V_0]^{DNA}$ values are distributed across a rather wide range of 0.99–1.14. When water molecules are included in the calculations the mean $[V/V_0]^{DNA}$ value drops to 1.01 and the range of individual values is narrower (0.97–1.06) (Fig. 2A). Two protein–DNA complexes have a $[V/V_0]^{DNA}$ value of 1.06: NFκB-p50 (1nfk) and the TATA box binding protein (1ais). These complexes also have a rather large $[V/V_0]^{PROT}$ value of 1.03, computed for the buried protein atoms in the interface. The higher ratios and wider distributions computed in the absence of water molecules are due to poorer statistics resulting from the smaller number of buried interface atoms, and from the fact that a fraction of the buried atoms is probably not optimally surrounded by neighbors (13,16).

Using our radii for the DNA atoms, we were also able to recompute the volume ratios for the protein portion of the interfaces $[V/V_0]^{PROT}$. This yielded a mean volume ratio of 1.02 and a range of 0.97–1.08, for individual values (Table 5 and Fig. 2B), in good agreement with the results obtained previously (22). We found only a moderate correlation between the V/V_0 values of the protein and DNA interface atoms, with a linear correlation coefficient of 0.6. For example, four of the complexes with $[V/V_0]^{DNA} < 1.0$ (1bpy, 1fjl, 1ign and 1lat) also have $[V/V_0]^{DNA}$ values ≤1.0. Similarly, the five complexes with the highest values of $[V/V_0]^{DNA}$ (1a3q, 1ais, 1cdw, 1nfk and 1rvc), also have relatively large $[V/V_0]^{PROT}$ values.

These results taken together indicate that, on average, both DNA and protein atoms buried in protein–DNA interfaces are

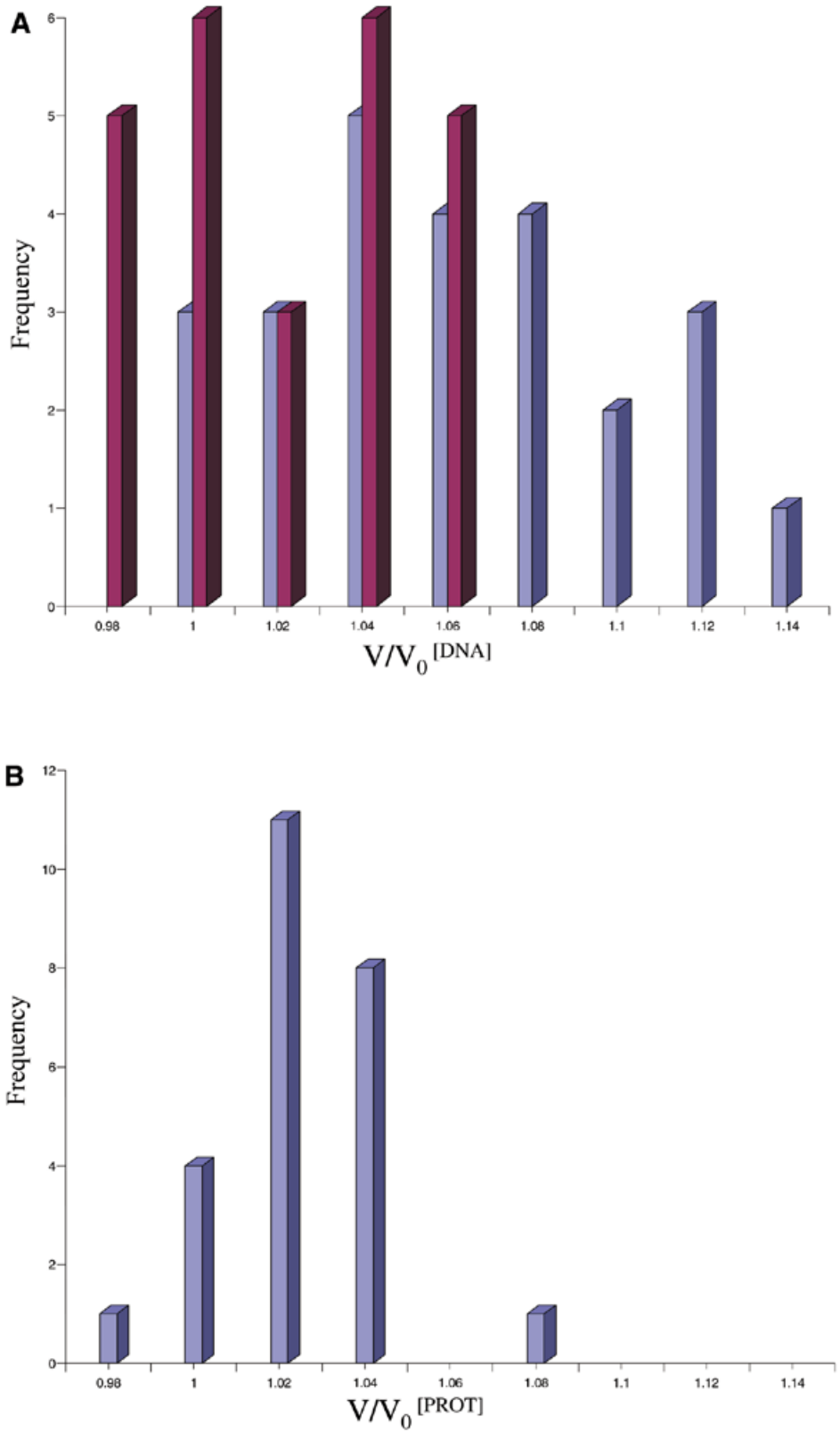

Figure 2. Histograms of the volume ratios of the DNA and protein atoms at the interfaces of the 25 protein–DNA complexes of Table 1. (**A**) $[V/V_0]^{DNA}$ values, computed in the presence (purple) and absence (light blue) of crystallographic water molecules. (**B**) $[V/V_0]^{PROT}$ values, computed in the presence of crystallographic water molecules. V is the sum of the volumes of nucleic acid or protein atoms buried at interfaces, V_0 is the sum of standard reference volumes for atoms buried inside B-DNA or the protein core. All volumes are computed using the Radical Planes method.

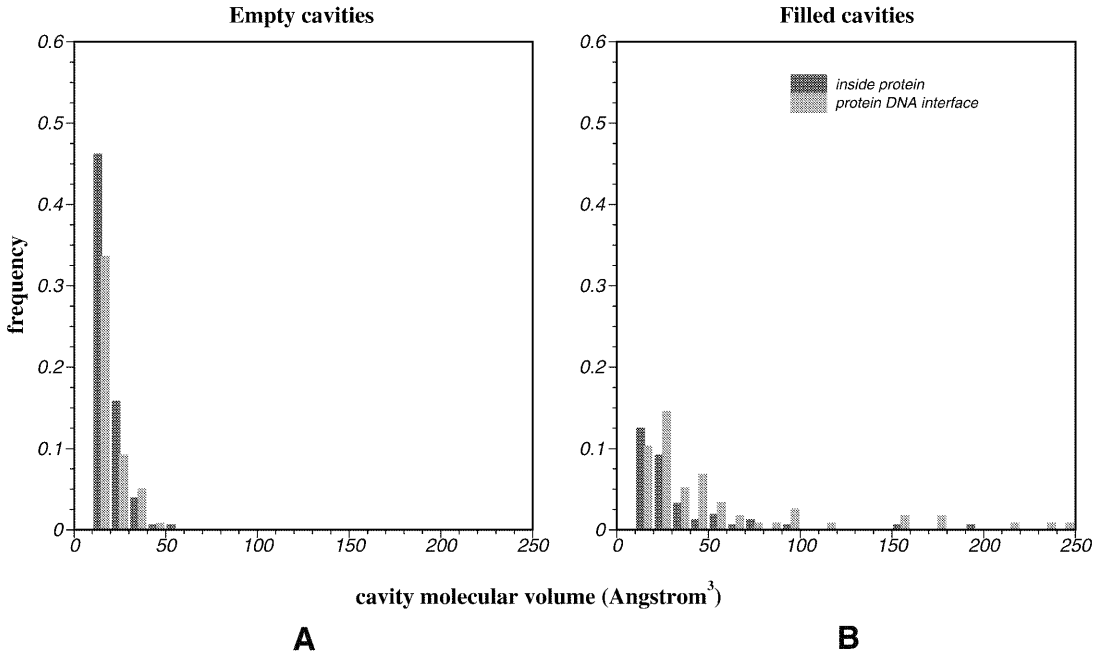

Figure 3. Histograms of the volumes of empty and water-filled cavities in protein–DNA interfaces and inside the protein components of protein–DNA complexes normalized to the total number of cavities in each of the two subsets. (**A**) The volumes of empty cavities in the interfaces (light gray) and inside the protein components (dark gray) of the 25 protein–DNA complexes analyzed in this study. (**B**) The volumes of filled cavities in the interfaces (light gray) and inside the protein components (dark gray) of the analyzed complexes.

nearly as close-packed as their counterparts in B-DNA, and inside the protein core, respectively. They furthermore confirm previous findings (22), that water molecules play an important role in fostering the tight packing at these interfaces.

Volume of cavities in protein–DNA interfaces. The packing at the interfaces in the protein–DNA complexes is further evaluated by computing the volume of packing defects, or cavities, in these interfaces. A cavity is defined here as a region of space, which contains no atomic positions assigned by the crystallographer and is completely surrounded by protein or DNA atoms, whose volume is delimited by the so-called molecular surface (9,33).

Cavity locations and volumes are computed using the software SURVOL (28) (see Materials and Methods). Two calculations are carried out. One in the absence of crystallographic water molecules and a second in their presence. Cavities identified in the first calculations but not in the second, are classified as 'water-filled', whereas the cavities identified in the presence of water molecules are classified as 'empty'. It should be clearly understood, however, that our calculations cannot distinguish between cavities representing true voids from those containing disordered water molecules or other disordered groups that are invisible in the electron density map.

Figure 3A shows the volume distributions of individual empty cavities, normalized to the total number of cavities, computed respectively, in the interfaces and inside the protein moieties of our sample of protein–DNA complexes. In comparison, Figure 3B displays the equivalent distributions for water-filled cavities (see Materials and Methods).

We see that the volume distribution of empty cavities in the analyzed interfaces is very similar to that in the protein interior, with however a lower fraction of small cavities (<30 Å^3) in the protein–DNA interfaces (~45%) than in the protein interior (~60%). On the other hand, we find that a larger proportion of the cavities in protein–DNA interfaces (~60%) is filled with water, than in the protein interior (30%). Water-filled cavities in the interfaces also tend to be larger, with a maximum volume of 657 Å^3 (Table 6), whereas inside the protein, the volume of the largest filled cavity is 190 Å^3.

Table 6 lists for each of the analyzed protein–DNA interfaces, the number of empty and water-filled cavities in the interface along-side the total, minimum and maximum volumes of each type of cavity. Of the total of 25 interfaces analyzed in our dataset, four contain no empty cavities. These are 1lmb (lambda repressor operator), 1tc3 (*Caenorhabditis elegans* transposase),1tsr (p53 core) and 2dgc (GCN4 ATF site). Interfaces of the latter two complexes contain no water-filled cavities either.

The total volume of empty cavities in individual interfaces displays appreciable variability. The smallest non-zero total volume (11.7 Å^3) is found for the DNA binding domain of the glucocorticoid receptor (1lat), which also features a $[V/V_0]^{DNA} < 1$. The largest total empty cavity volume (695.4 Å^3) is observed in the structure of the TRP repressor/operator half-tandem complex (1trr). The majority of this volume (636.1 Å^3) belongs to a very large elongated cavity (not plotted in Fig. 3 for clarity), located between the two monomers of the repressor and reaching into the protein–DNA interface. Visual inspection suggests that very small atomic displacements under thermal motion would probably turn this cavity into a channel communicating with bulk solvent. It is thus very likely

Table 6. Volumes of empty and water-filled cavities in protein–DNA interfaces

PDB code	Complex	Empty cavities				Filled cavities			
		no	Molecular volume			no	Molecular volume		
			Total	Minimum	Maximum		Total	Minimum	Maximum
1bpy	DNA polymerase	3	41.5	12.1	14.8	3	250.5	17.6	211.9
1t7p	Phage T7	3	56.9	12.0	28.7	6	569.1	22.7	179.1
1dnk	Dnase I	1	30.5	30.5	30.5	0	0.0	-	-
1rvc	EcoRV	3	73.1	11.6	49.7	5	569.2	21.6	240.1
1hcr	Hin recombinase	1	15.9	15.9	15.9	0	0.0	-	-
1tc3	Transposase	0	0.0	-	-	1	12.0	12.0	12.0
1ign	RAP1 telomere	1	12.6	12.6	12.6	2	135.3	41.6	93.7
1lmb	Lambda repressor	0	0.0	-	-	3	69.2	20.3	23.5
1tro	Trp repressor	6	113.7	12.3	24.8	2	39.7	18.5	21.2
1trr	Trp repressor,half site	5	695.4	12.5	636.1	3	710.0	21.4	657.9
1fjl	Paired dimer	2	49.0	17.7	31.3	5	185.7	15.1	79.7
1pue	Pu1-ETS domain	1	12.8	12.8	12.8	1	26.2	26.2	26.2
2dgc	GCN4,ATF site	0	0.0	-	-	0	0.0	-	-
1aay	Zif268	2	51.2	23.2	28.0	2	142.3	48.8	93.5
1mey	Designed	4	76.9	12.9	32.4	1	155.2	155.2	155.2
1hcq	Estrogen receptor	1	11.8	11.8	11.8	7	241.7	16.3	62.7
1lat	Glucocortiocoid	1	11.7	11.7	11.7	0	0.0	-	-
2nll	Retinoid receptor	3	73.3	19.4	29.8	7	210.1	14.7	52.3
1ais	TBP-TFIIB	8	135.3	12.5	25.3	1	87.8	87.8	87.8
1cdw	TBP, human	4	59.0	12.0	18.1	0	0.0	-	-
1a3q	NFk-B p52	1	33.4	33.4	33.4	0	0.0	-	-
1nfk	NFk-B p50	2	74.4	36.2	38.2	1	69.5	69.5	69.5
1tsr	p53 core	0	0.0	-	-	0	0.0	-	-
2bop	E2 domain	2	29.2	14.6	14.6	2	42.1	21.0	21.0

that it contains disordered solvent molecules, which are not visible in the electron density map.

The remaining empty space is distributed amongst small cavities. Two other interfaces display a large total empty volume: those of 1ais (TBP, TFIIB) and 1tro (TRP repressor dimer). But here this volume is distributed amongst small cavities of ≤30 Å^3.

Table 6 shows that the total volume of water-filled cavities in interfaces of individual complexes displays even wider variability than the volume of empty cavities. Interesting cases are the structures of the DNA polymerase complex with gapped DNA (1bpy), the *Eco*RV–DNA complex (1rvc), the T7 DNA polymerase–thioredoxin complex (1t7p), the estrogen receptor–DNA complex (1hcq), the retinoid receptor–DNA complex (2nll) and once again the TRP repressor/operator half-tandem complex (1trr). All these complexes have three or more filled cavities with a volume superior to 200 A^3 at the protein–DNA interface.

A pictorial representation of two of these complexes, *Eco*RV–DNA and the retinoic acid receptor–DNA complex, is given in Figure 4. In the first complex with the endonuclease (Fig. 4A), the water-filled cavities are clustered at the active site of the enzyme, and are centrally located within the protein–DNA interface. In the second complex (Fig. 4B) the filled cavities are more evenly distributed across the interface, especially in one of the monomers. In both cases the buried water molecules play a key role as building blocks of the molecular interfaces.

Several interesting conclusions can be drawn from this analysis. We see that protein–DNA interfaces contain, on average, a somewhat smaller volume of the empty space than in the interior of DNA binding proteins. On the other hand, they harbor more water-filled pockets than in the core of these proteins, in line with the polar character of the atoms in these interfaces. These filled pockets contain buried water molecules, which form H-bonds with the DNA and protein atoms and with one another, as will be described elsewhere (Tomás-Oliveira,I., Nadassy,K., Alberts,I., Janiu,J. and Wodak,S.J., manuscript in preparation). These buried waters seem to be an integral part of the molecular interfaces and therefore play a key role in fostering close packing at these interfaces. It is therefore not surprising that they also play an important role in fostering specific protein–DNA recognition, as already suggested (34).

DISCUSSION

Atomic volumes in DNA and at interfaces

This study presents the first calculation of the volumes occupied by atoms and residues in double-stranded DNA. Mean atomic volumes and standard deviations were computed from several sets of structures. From high-resolution structures of B-DNA, A-DNA and Z-DNA in the NDB and from a set of structures in the CSD, extracted by searching for bases, sugar and phosphate moieties.

As expected, the trends in the mean atomic volumes from all the sets are dictated by the chemical type of the atomic group and its covalent bonding environment.

The mean atomic volumes in double helical DNA structures from the NDB were found to be ~5% smaller than those in the CSD structures, indicating that DNA structures are more closely packed than crystals of related nucleic acid molecules. This clearly arises from the presence of hydrogen bonding and stacking interactions in DNA, whereas in the CSD structures,

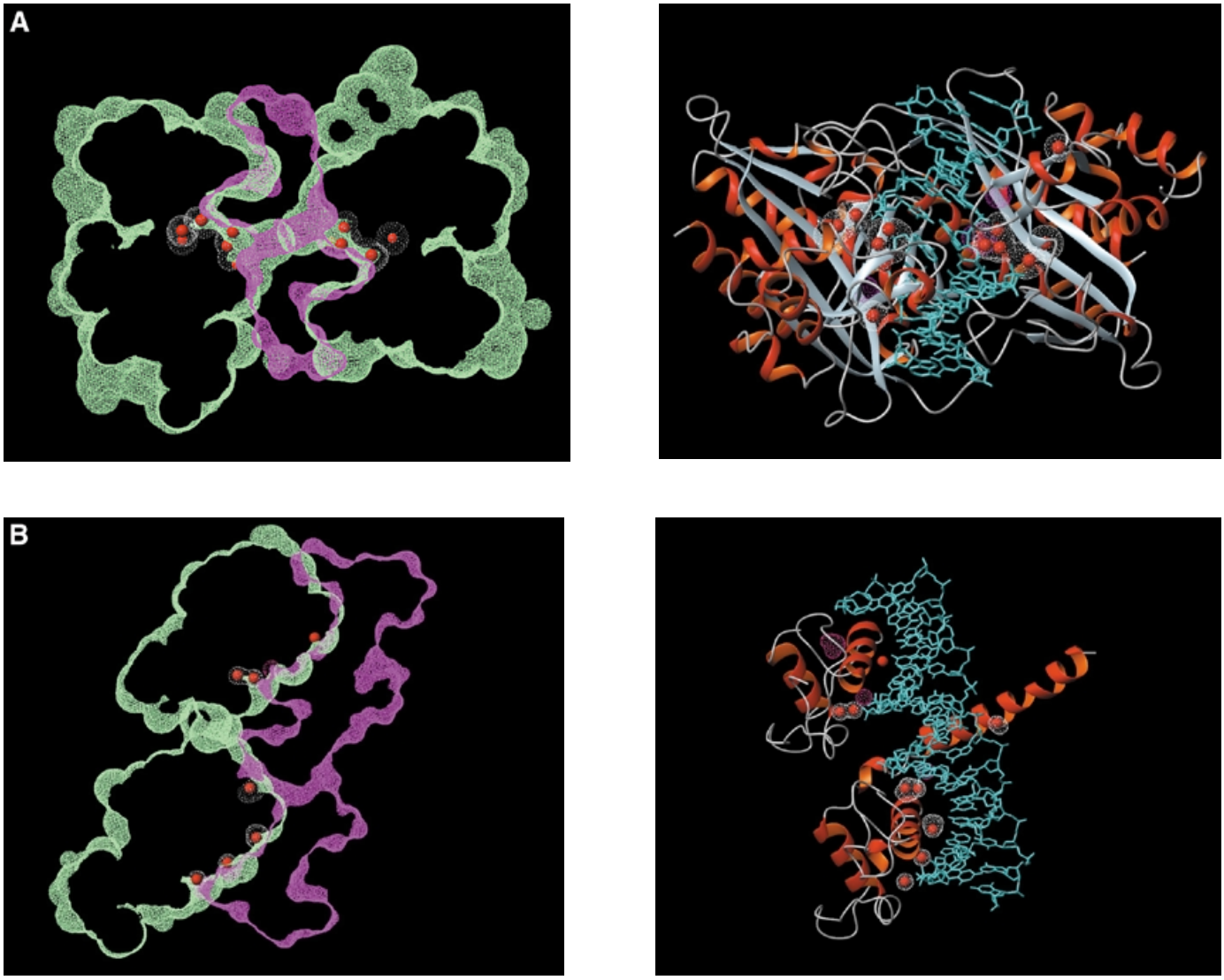

Figure 4. Pictorial illustrations of two protein–DNA complexes with large volumes of water-filled cavities in their interfaces. This figure illustrates the cases of two complexes from our sample, which feature amongst the largest volumes of water-filled cavities in their interface. (**A**) *Eco*RV–DNA complex (PDB code 1rvc). The left-hand side shows a slice through the surfaces of the protein (two monomer) and DNA (two segments, 11 bp long) portions of the complex in the crystal structure. The DNA molecular surface is in magenta, and the protein accessible surface is in light green. The accessible surface area is depicted for the protein in order to improve the display of the cavities, whose surface is shown in white dots. Because the protein accessible surface is obtained by rolling a 1.4 Å sphere over the van der Waals surface of the protein, the corresponding surface contours are seen to intersect in places with the molecular surface of the DNA. The right-hand side displays the *Eco*RV–DNA complex in a similar orientation as on the left, but highlighting the secondary structure elements of the protein. α Helices are shown as flat red–orange ribbons, β strands as gray ribbons and loops as thin gray strings. The DNA moiety is displayed as a turquoise stick model. Buried water molecules are depicted as filled red spheres. They are located in a cluster of cavities, at the enzyme active site near the center of the protein–DNA interface. (**B**) The retinoic acid receptor–DNA complex (PDB code 2nll). The left-hand side shows a slice through the surfaces of the protein (two monomers) and DNA (two segments, 18 bp long) portions of the complex, with the DNA and protein surfaces computed as detailed above and displayed using the same color convention as in (A). The right-hand side displays the same complex in the same projection, but somewhat different orientation as on the left, but highlighting the secondary structure elements of the protein [see (A) for details]. Water-filled cavities are displayed by their molecular surface (white dots), and the crystallographic waters molecules (filled red spheres) which they contain. Two empty cavities (delimited by their molecular surface in magenta) are visible at the bottom of the right-hand picture. At the bottom we also see a buried water molecule whose surrounding cavity is not visible, being beyond the size detectable by the program.

where nucleic acid groups occur in very diverse environments, such interactions are often absent. It is interesting that this trend parallels that observed for the atomic volumes in the protein interior versus those in amino acid crystals (8,15). Thus, the two major types of macromolecules in living cells, DNA and proteins, representing highly organized systems with many specific interactions, presumably perfected through evolution, are also more compact on average than crystals of their building blocks, obtained in the laboratory, where intermolecular interactions are less specific and more diverse.

Another finding of our study is that with the exception of a few atom types, the atomic volumes in A-DNA, B-DNA and Z-DNA are rather similar. But the results obtained for atoms in Z-DNA are based on a very small number of observations (particularly for the A and T bases), and need confirmation by further analysis.

In the second part of this work, we use the atomic volumes of B-DNA as a standard, to evaluate the packing efficiency of nucleic acid atoms buried in the interfaces of 25 high resolution protein–DNA complexes. This packing efficiency is expressed as the ratio of the atomic volume in the interface to the standard volumes. This ratio is found to be close to unity, on average, indicating that the DNA atoms in protein–DNA interfaces are as closely packed as in B-DNA. This taken together with our previous findings about protein atoms in these interfaces being as closely packed as atoms inside proteins (22), leads us to conclude that the packing efficiency in protein–DNA interfaces is as high as in the respective macromolecules.

An important factor in our calculations is the contribution from the crystallographic water molecules in the interfaces. The proportion of DNA atoms buried in the interfaces increases from 30 to 54% and their packing efficiency improves, when these molecules are included in the calculation. Similar observations were made for the packing of protein atoms in these interfaces (22), as well as in protein–protein interfaces (21).

This general picture is further supported by our analysis of empty and water-filled cavities in the protein–DNA interfaces. The frequency and volumes of empty cavities are somewhat below those found in the protein interior. But those of solvent-filled cavities, is clearly higher. This indicates that protein–DNA interfaces are more hydrated than the protein interiors and emphasizes the importance of solvent molecules in enhancing packing at these interfaces.

Table 7. Shape complementarity at protein–DNA interfaces

PDB code	Complex	$\overline{V/V_0}$ [a]	Sc [b]	Gap Index [c]
1bpy	DNA polymerase	0.99	0.69	3.76
1t7p	Phage T7	1.04	0.59	3.78
1bhm	BamHI	1.01	0.66	2.22
1dnk	Dnase I	1.02	0.71	2.34
1rvc	EcoRV	1.04	0.68	1.89
1hcr	Hin recombinase	1.00	0.72	0.80
1tc3	Transposase	0.99	0.75	2.29
1ign	RAP1 telomere	0.98	0.74	1.51
1lmb	Lambda repressor	1.03	0.62	2.50
1tro	Trp repressor	1.03	0.59	1.94
1trr	Trp repressor,half site	1.00	0.64	1.58
1fjl	Paired dimer	0.98	0.71	2.50
1pue	Pu1-ETS domain	1.00	0.68	2.08
2dgc	GCN4,ATF site	1.01	0.58	3.14
1aay	Zif268	1.01	0.66	1.60
1mey	Designed	1.01	0.64	1.99
1hcq	Estrogen receptor	1.01	0.56	0.91
1lat	Glucocortiocoid	1.00	0.65	3.00
2nll	Retinoid receptor	1.03	0.64	2.50
1ais	TBP-TFIIB	1.05	0.58	2.92
1cdw	TBP, human	1.04	0.64	2.30
1a3q	NFk-B p52	1.04	0.71	3.61
1nfk	NFk-B p50	1.04	0.67	3.00
1tsr	p53 core	1.06	0.60	4.33
2bop	E2 domain	1.02	0.67	1.39

[a] $\overline{V/V_0} = ([V/V_0]^{PROT} + [V/V_0]^{DNA})/2$.
[b]Shape correlation index of Laurence and Colman (25) calculated using the program space.
[c]Gap index of Jones and Thornton (26) calculated using the authors' internet resource (http://www.biochem.ucl.ac.uk/bsm/DNA/server/).

Shape complementarity

Two estimates of shape complementarity have previously been used to evaluate the extent of packing in interacting macromolecules. One is the so-called 'gap volume index' of Laskowski (31), used by Jones and Thornton (26) and Jones *et al.* (27). This index is computed as the volume available between the solvent accessible surfaces of the molecular components of the complex, divided by the interface area. (see http://www.biochem.ucl.ac.uk/bsm/DNA/server/parameter_def.html). The other is the shape correlation index (S_c) of Lawrence and Colman (25), derived from the distance between points on the surfaces of the interacting molecules and the angle between the vectors normal to these surfaces. The S_c index was used to evaluate protein–protein interfaces, and the gap volume index was used to evaluate protein–protein (26) and protein–DNA interfaces (27).

It seemed of interest to compare our volume based evaluations of the interfaces with these criteria. Since our calculations treated the protein and DNA portions of the interfaces independently, we computed a volume ratio describing each interface as a whole, defined as the average of the volume ratios for the protein and DNA atoms. These ratios, together with the values for the gap volume and S_c indices, are listed in Table 7. Better shape complementarity should appear as a lower value of the volume ratio and the gap volume index, and a higher value of S_c. In agreement with the analysis of Jones *et al.* (27), we observe that the computed gap volume indices span a very large spectrum of values ranging from ~0.8–4.3 Å, whereas the S_c values span a narrower range, similar to that found in protein–protein interfaces.

We find a very poor correlation between the $\overline{V/V_0}$ values, or for that matter also the volume ratios of the protein and DNA components, with either of the shape complementarity indices. The linear correlation coefficients were 0.4 and 0.5 with the gap volume and S_c indices, respectively; in general, however, when $\overline{V/V_0}$ indicates rather poor packing relative to the interior of protein and DNA molecules, the gap volume index and S_c values are in rough agreement. For example, out of the seven values of ≥1.04, for PDB entries 1a3q, 1ais, 1cdw, 1nfk, 1rcv, 1tp7 and 1tsr, five are associated with relatively high gap volume indices of >2.9. Six of these entries, with the exception of 1a3q, also have S_c values <0.68, which is indicative of relatively poor shape complementarity.

Similarly, well packed complexes with $\overline{V/V_0} \approx 1$ (PDB codes, 1hcr, 1ign and 1tc3), have relatively low gap volume indices and high S_c values. But many counter examples to this rule are also observed.

The poor correlation between our volume ratios and the two shape complementarity indices may be explained by the inherent difference between the properties measured by volume calculation and by the two shape complementarity indices. We measure the volume occupied only by atoms that are completely surrounded by other atoms in the interfaces. Hence, the volumes of atoms lining the exterior surface, and those lining empty internal cavities are not computed. Also, our volume calculations say nothing about how patchy an interface is and evaluate only the packing in the patches where direct, or water mediated, intermolecular interactions form.

In contrast, the gap volume and S_c indices measure geometric properties of the interacting surfaces. The S_c index evaluates the correlation between the shapes of the interacting surfaces, whereas the gap volume index aims at measuring how patchy an interface is.

Thus, not only is there no obvious relation between these two geometric properties and the packing efficiency calculated here but also between the two geometric properties themselves. It is hence not too surprising that the corresponding indices not only display poor correlation with our volume ratios, but also with each other. We find indeed that linear correlation coefficient between the S_c and gap volume index is only 0.2.

These observations furthermore explain why Jones *et al.* (27) reach different conclusions about the packing in protein–DNA interfaces, evaluated on the basis of the gap volume index, than in the present study. They interpret the wide range in gap volume indices to mean that the packing efficiency of protein–DNA interfaces varies significantly. They suggest, for example, that monomeric proteins have more tightly packed protein–DNA interfaces than dimeric proteins, with a tendency for DNA bound enzymes to display more segmented protein–DNA interfaces than the complexes with transcription factors.

Our volume ratios reveal no such differences. But there remained the possibility that the volume of empty cavities in the interfaces, computed in the absence of water molecules may represent a measure similar to the gap volume, also computed in the absence of water molecules. This was checked here by computing the cavity volume index for our interfaces, defined as the total cavity volume divided by the total area buried at each interface, and comparing this index with the gap volume index. However, this comparison also revealed a poor correlation between these two indices (linear correlation coefficient of 0.2). This confirms that the gap volume of Thornton and coworkers (26,27) and Laskowski (31) is not simply related to the cavity volumes computed here, and hence shows that it does not represent the volume of the empty 'holes' formed when two poorly complementary interfaces interact. It seems likely, on the other hand, that the gap volume index is more representative of the complementarity between the surface regions at the periphery of the interface, which are accessible to bulk solvent in the complex. Indeed, since spheres of diameter as large as 10 Å are used in the gap volume calculations (see Materials and Methods), those with such diameter would tend to be in contact with the surfaces of atoms in the peripheral regions, rather then fill cavities located within the interface proper and would hence contribute more to the gap volume. The relevance of measuring the surface complementarity in these peripheral regions, however, remains to be demonstrated.

CONCLUSIONS

Protein–DNA recognition in particular, and protein–nucleic acid recognition in general, play a central role in biology. Our study has for the first time evaluated atomic volumes in DNA crystal structures and compared them to those of nucleic acid groups in small molecule crystals. This comparison confirms the compact nature of B-DNA. Our analysis of the atomic volumes of the DNA portion in protein–DNA interfaces, and of the empty and water-filled cavities at these interfaces, show that these interfaces are significantly more hydrated than the protein interior and that water-hydration plays a key role by fostering close packing, and therefore also in specific recognition.

Analyses such as these can be readily applied to the growing number of protein–DNA and protein–RNA complexes, solved at high resolution, and should provide valuable insights into the principles that govern recognition in these important systems.

ACKNOWLEDGEMENTS

Our thanks go to the computer systems group of the European Bioinformatics Institute, and to Jean Richelle at the SCMBB, for their valuable help. We are also deeply indebted to Koji Ogata for valuable help with generating the color figures. K.N. gratefully acknowledges support from the University of Stirling, and I.T.-O. thanks the Fundação para a Ciência e Tecnologia, Portugal (Praxis XXI/BD/5697/95) and the Free University of Brussels for support. This work was also supported by the Action de Recherches concertées de la Communauté Française de Belgique, project no. 97/01-211.

REFERENCES

1. Murphy,K.P. and Gill,S.J. (1991) Solid models compounds and the thermodynamics of protein unfolding. *J. Mol. Biol.* **222**, 699–709.
2. Richards,F.M. (1997) Protein stability: still an unsolved problem. *Cell. Mol. Life Sci.*, **53**, 790–802.
3. Ptitsyn,O.B., Pain,R.H., Semisotnov,G.V., Zerovnik,E. and Razgulyaev,O.I. (1990) Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett.*, **262**, 20–24.
4. Hughson,F.M., Barrick,D. and Baldwin,R.L. (1991) Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. *Biochemistry*, **30**, 4113–4118.
5. Ericksson,A.E., Baase,W.A., Zhang,X.J., Heinz,D.W., Blaber,M., Baldwin,E.P. and Matthews,B.W. (1992) Response of a protein structure to cavity-creating mutations and its relations to the hydrophobic effect. *Science*. **255**, 178–183.
6. Voronoi,G.F. (1908) Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J. Reine Angew. Math.*, **134**, 198–287.
7. Gellatly,B.J. and Finney,J.L. (1982) Calculation of protein volumes: an alternative to the Voronoi procedure. *J. Mol. Biol.*, **161**, 305–322.
8. Richards,F.M. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.*, **82**, 1–14.
9. Richards,F.M. (1977) Areas, volumes, packing and protein structures. *Annu. Rev. Biophys. Bioeng.*, **6**, 151–176.
10. Bondi,A. (1964) van der Waals volumes and radii. *J. Phys. Chem.*, **68**, 441–451.
11. Chothia,C. and Janin,J. (1975) Principles of protein-protein recognition. *Nature*, **256**, 705–708.
12. Li,A.J. and Nussinov,R. (1998) A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins*, **32**, 111–127.
13. Tsai,J., Taylor,R., Chothia,C. and Gerstein,M. (1999) The packing density in proteins: standard radii and volumes. *J. Mol. Biol.*, **290**, 253–266.
14. Finney,J.L. (1975) Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J. Mol. Biol.*, **96**, 721–732.
15. Harpaz,Y., Gerstein,M. and Chothia,C. (1994) Volume changes on protein folding. *Structure*, **2**, 641–649.
16. Pontius,J., Richelle,J. and Wodak,S.J. (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, **264**, 121–136.
17. Fleming,P.J. and Richards,F.M. (2000) Protein packing: dependence on protein size, secondary structure and amino acid composition. *J. Mol. Biol.*, **299**, 487–498.
18. Pontius,J. (1997) Atomic volumes in protein crystallographic structures and their use in structure validation. Université Libre de Bruxelles, Brussels.

19. Gerstein,M. and Chothia,C. (1996) Packing at the protein-water interface. *Proc. Natl Acad. Sci. USA*, **93**, 10167–10172.
20. Gerstein,M., Tsai,J. and Levitt,M. (1995) The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.*, **249**, 955–966.
21. Conte,L.L., Chothia,C. and Janin,J. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.
22. Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
23. Berman,H.M., Olson,W.K., Beveridge,D.L. Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
24. Allen,F.H., Bellard,S., Brice,M.D., Cartwright,B.A., Doubleday,A., Higgs,H., Hummelink,T., Hummelink-Peters,B.G., Kennard,O., Motherwell,W.D.S., Rodgers,J.R. and Watson,D.G. (1979) The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr. B*, **35**, 2331–2339.
25. Lawrence,M.C. and Colman,P.M. (1993) Shape complementarity at protein/protein interfaces. *J. Mol. Biol.*, **234**, 946–950.
26. Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
27. Jones,S., van Heyningen,P., Berman,H.M., and Thornton,J.M. (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–996.
28. Alard,P. (1991) Calcul de surface et d'énergie dans le domaine des macromolécules, Université Libre de Bruxelles, Brussels.
29. Connolly,M.L. (1983) Analytical molecular surface calculation. *J. Appl. Crystallogr.*, **16**, 548–558.
30. Alard,P. and Wodak,S.J. (1991) Detection of cavities in a set of interpenetrating spheres. *J. Comp. Chem.*, **12**, 918–922.
31. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.
32. Fletcher,D.A., McMeeking,R.F. and Parkin,D.J. (1996) The United Kingdom Chemical Database Service. *J. Chem. Inf. Comput. Sci.*, **36**, 746–749.
33. Connolly,M.L. (1985) Atomic size packing defects in proteins. *Intl J. Pept. Protein Res.*, **28**, 360–363.
34. Janin,J. (1999) Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure Fold Des.*, **7**, R277–R279.