Article No. jmbi.1999.2829 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 290, 253–266

JMB

The Packing Density in Proteins: Standard Radii and Volumes

Jerry Tsai[1], Robin Taylor[2], Cyrus Chothia[3] and Mark Gerstein[4]

[1]*Department of Biochemistry University of Washington Box 357350, Seattle WA 98195, USA jotter@felix.bchem.washington.edu*

[2]*Cambridge Crystallographic Data Centre, 12 Union Road Cambridge, CB2 1EZ, England taylor@ccdc.cam.ac.uk*

[3]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England chcl@mrc-lmb.cam.ac.uk*

[4]*Department of Molecular Biophysics and Biochemistry Yale University, Bass Center 266 Whitney Avenue P.O. Box 208114, New Haven CT 06520-8114, USA mark.gerstein@yale.edu*

The sizes of atomic groups are a fundamental aspect of protein structure. They are usually expressed in terms of standard sets of radii for atomic groups and of volumes for both these groups and whole residues. Atomic groups, which subsume a heavy-atom and its covalently attached hydrogen atoms into one moiety, are used because the positions of hydrogen atoms in protein structures are generally not known. We have calculated new values for the radii of atomic groups and for the volumes of atomic groups. These values should prove useful in the analysis of protein packing, protein recognition and ligand design. Our radii for atomic groups were derived from intermolecular distance calculations on a large number (~30,000) of crystal structures of small organic compounds that contain the same atomic groups to those found in proteins. Our radii show significant differences to previously reported values. We also use this new radii set to determine the packing efficiency in different regions of the protein interior. This analysis shows that, if the surface water molecules are included in the calculations, the overall packing efficiency throughout the protein interior is high and fairly uniform. However, if the water structure is removed, the packing efficiency in peripheral regions of the protein interior is underestimated, by ~3.5%.

Keywords: atomic group radii and volumes; Voronoi polyhedra; packing efficiency; protein interior

Introduction

Previous work has demonstrated that, in general, residues in the interior of proteins are closely packed with mean volumes a little smaller than those they have in amino acid crystals (Richards, 1974; Harpaz *et al*., 1994; Chothia, 1975). This observation, however, leaves open the question of whether or not the packing density is higher in certain parts of proteins, such as the deeply buried interior, than it is in other regions, such as those on the periphery. Here, we report an investigation that answers this question. As a consequence of this study, we also developed a set of standard values for the radii and volumes of atomic groups in proteins. These should be of general use and help, for example, in calculations relating to analysis of protein structures, protein recognition and ligand design (Richards, 1979; Shoichet & Kuntz, 1991, Kocher *et al*., 1996; Peters *et al*., 1996; Liang *et al*., 1998; Lo Conte *et al*., 1999; Nadassy *et al*., 1999).

The most rigorous method for calculating the packing density in protein structures involves the construction of Voronoi polyhedra around atomic groups. The proper application of this method requires accurate values for the van der Waals (VDW) radii of the various atomic groups. The VDW radii of individual atoms have been well documented (Bondi, 1964; Rowland & Taylor, 1996). In proteins, however, the position of hydrogen atoms is not generally known. This has meant that hydrogen atoms are subsumed into the ''heavy'' atoms to which they are covalently linked, creating atomic groups. Thus, radii for atomic groups, such as the methyl group (-CH3), apply to the groups as a whole. Several sets of radii for atomic groups are available the literature, but there are appreciable differences among them (see below). We therefore determined a new set of radii, based on an extensive and detailed study of

Abbreviations used: VDW, van der Waals; ProtOr, protein organic.

0022-2836/99/260253–14 $30.00/0

the intermolecular contacts made by atomic groups in small molecule crystals with the same constituents as proteins.

The VDW and hydrogen bond radii of atomic groups in proteins

Nomenclature

Here, the atomic groups found in proteins are given labels of the general form XnHm, where X indicates the chemical nature of the non-hydrogen atoms; n, their valence; and Hm, the number (m) of hydrogen atoms attached to the non-hydrogen atom. For instance, C3H1 is a trigonal carbon atom with one attached hydrogen. Table 1 gives a list of the 13 different atomic groups that are treated in this study.

Procedure for determining the radii of atomic groups

To determine VDW radii of these atomic groups, we used small molecule organic structures from the Cambridge Structural Database (CSD, Allen *et al*., 1991). This database was chosen because the amount and precision of its data are considerably higher than that available from the crystal structures of proteins. The methodology used here to determine the radii of these groups is described in detail in a previous study by Rowland & Taylor (1996). Its use here involved the following steps. First, a suitable subset of the CSD (October 1996 version) was chosen. This was comprised of database entries satisfying the following criteria: (1) *R*-factor less than 10%; (2) no disorder; (3) not an ''error set,'' as defined in the CSD; (4) not a polymer; and (5) no elements present other than C, H, N, O, S. When more than one determination of the same compound was present in the CSD, only one (chosen arbitrarily) was used. The subset comprised 30,111 structures in all.

The program QUEST (Rowland & Taylor, 1996) was then used to find intermolecular contacts between each pair of the 13 groups listed in Table 1, except thiol (S2H1) which is very rare in the CSD. This meant that we collected data on the contacts formed between 66 ($12 \times 11/2$) pairs of different groups and 12 pairs, where the contacts are between groups of the same kind: 78 in all. For a given pair, all contacts were tabulated out to a distance of $V + 1.5$ Å, where V is the sum of the Bondi VDW radii (Bondi, 1964) of the non-hydrogen atoms involved in the contact. Each resulting non-bonded distance distribution was plotted as a histogram. An empirical fitting procedure was then used to estimate d, the distance at which the histogram reaches half its maximum height. (It was demonstrated by Rowland & Taylor (1996) that the half-height value can be estimated more precisely than the distance at which the distribution reaches its maximum, and that the resulting d values are almost identical with the sum of the corresponding Bondi VDW radii.)

A total of 38 of the 78 histograms were rejected at this stage, either because they contained insufficient data or because the shape of the histogram did not permit its half-height position to be estimated reliably. Of the remaining 40, 32 involved VDW contacts and were used in the determination of VDW radii. The other eight histograms involved hydrogen-bonding contacts, e.g. O2H1$\cdots$O1H0 and were used to determine ''hydrogen bond radii''.

A least-squares fitting procedure was then used to generate a set of radii, $r(i)$, which minimised the function:

$$f = \sum_{i=1}^{n} \sum_{j=1}^{n} w(ij) \times \{\mathrm{d}(ij) - [r(i) + r(j)]\}^2$$

where the outer summation is over all the groups in Table 1, excluding thiol. Here d(ij) is the half-height value for the non-bonded distance distribution between groups i and j. When d(ij) could not be estimated or corresponds to a hydrogen-bond contact, w(ij) is zero. Otherwise, w(ij) is unity.

This procedure resulted in an initial set of radii. We compared these with the radii derived previously by Chothia (1975); see the second column of Table 2 below. We found good agreement between the two radii sets for carbon, oxygen and

Table 1. Atomic groups in proteins

Atomic group	Chemical formula	Valency of X atom (n)	Hydrogen atoms bonded to X (m)	Non-hydrogen atoms bonded to X
N3H0	>N-	3	0	3
N3H1	>NH	3	1	2
N3H2	-NH_2	3	2	1
N4H3	-NH_3^+	4	3	1
O1H0	=O or -O^-	1	0	1
O2H1	-OH	2	1	1
C3H0	-C<	3	0	3
C3H1	-CH-	3	1	1
C4H1	-CH<	4	1	3
C4H2	-CH_2-	4	2	2
C4H3	-CH_3	4	3	1
S2H0	-S-	2	0	2
S2H1	-SH	2	1	1

Table 2. Sets of VDW radii for atomic groups in proteins

Atomic group	ProtOr Radii[a]	Chothia[b] Radii[a]	Richards[c] Radii[a]	Li & Nussinov[d] Radii[a]
C3H0	1.61	1.76	1.70	1.74-1.81
C3H1	1.76	1.76	1.70	1.74-1.82
C4H1	1.88	1.87	2.00	2.01
C4H2	1.88	1.87	2.00	1.88-1.92
C4H3	1.88	1.87	2.00	1.93
N3H0	1.64	1.50	1.70	1.65
N3H1	1.64	1.65	1.70	1.66-1.70
N3H2	1.64	1.65	1.60	1.62-1.67
N4H3	1.64	1.50	2.00	1.67
O1H0	1.42	1.40	1.40	1.49-1.52
O2H1	1.46	1.40	1.60	1.54
S2H0	1.77	1.85	1.80	1.94
S2H1	1.77	1.85	-	1.88

[a] Radii in Å.
[b] Chothia (1975).
[c] Richards (1974).
[d] Li & Nussinov (1998).

sulphur atoms, but discrepancies of up to 0.2 Å for the various types of nitrogen atoms. In the case of N4H3, the reason is clear. This group is always fully hydrogen-bonded in crystal structures. The N4H3 atom will therefore be prevented, sterically, from making anything other than long contacts to atoms which do not accept hydrogen bonds, as these atoms will necessarily be in the second non-bonded co-ordination sphere. Hence, the N4H3 VDW radius is inflated because it is based almost entirely on contacts to atoms in the second non-bonded co-ordination sphere. Similar effects, albeit less serious, are likely to occur for other types of atoms.

A second set of radii was therefore derived, using the same methodology as above but excluding from the analysis any contact that was not a nearest-neighbour interaction. This was done by checking that, for each A···B contact in the analysis, neither A nor B formed a shorter intermolecular contact (relative to the appropriate Bondi radii) to any other atom.

The N4H3 group, which the first calculation had shown not to make VDW contacts, was not included in the second calculation. This meant that data were collected on 66 atom pair combinations. Of these, six involved hydrogen-bond contacts and 12 gave histograms that had insufficient observations or too poor a shape to allow d values to be estimated. This left 48 which could be used to determine the VDW radii. The reason that comparatively few histograms were rejected in the second analysis was that the restriction to nearest-neighbour contacts greatly improved the shape of the distributions and allowed d to be estimated much more robustly.

The VDW radii given by this second calculation are listed in the first column of Table 2 and the hydrogen bond radii in Table 3. We will refer to this set of radii as the ProtOr (Protein Organic) radii set.

The values of these radii are well determined. The magnitude of the residual in the function given above is 0.184 Å^2. This corresponds to a root-mean-square difference between measured $d(ij)$ values and the sum of the radii $r(i) + r(j)$ of 0.06 Å. In fact, this value would be much lower were it not for a few individual large discrepancies, particularly O1H0···O1H0 (+0.25 Å) and N3H0···N3H0 (+0.13 Å). These large differences are entirely predictable: because these pairs of atoms have appreciable partial charges of the same sign, they repel each other and, therefore, do not approach each other quite as closely as their VDW radii would suggest.

The radii of atomic groups

The values obtained for the VDW radii of the three types of aliphatic carbon (Table 2) are identical to two decimal places (1.88 Å), indicating the numerical robustness of our method. Interestingly, the radius of C3H0 has decreased by almost 0.2 Å relative to the value found in the earlier set, presumably indicating that this atom can form short contacts in a direction approximately perpendicular to the plane of the trigonal carbon (see also below). The values for the three types of trigonal nitrogen atoms varied slightly: 1.64, 1.63 and 1.66 Å for N3H0, N3H1, and N3H2, respectively.

Table 3. ProtOr radii for hydrogen bonds

Atomic group	ProtOr Radii[a]
N3H1	1.48
N3H2	1.54
N4H3	1.28
O1H0	1.28
O2H1	1.32

[a] Radii in Å.

We regard these differences as insignificant and have therefore assigned an average value of 1.64 Å to all three groups. As described above, no VDW radius could be obtained for N4H3, reflecting the fact that atoms in the first co-ordination sphere of this group are invariably hydrogen-bonded. A VDW radius of N4H3 was, therefore, set to 1.64 Å, the same as for the trigonal nitrogen atoms. In addition, the radius of S2H1 (thiol) was set to 1.77 Å (i.e. identical with S2H0). This is the simplest reasonable decision that can be made in the absence of sufficient data for thiol.

''Hydrogen bond radii'' were also calculated for donor and acceptor groups from the histograms of distances between O1H0, O2H1, N3H1, N3H2 and N4H3 atomic groups (Table 3). The values obtained were 1.28, 1.32, 1.48, 1.54, and 1.43 Å, respectively.

Comparison of the radii described here and those given by previous calculations

In Table 2, where we list the ProtOr VDW radii, we also give the values determined by Richards (1974), Chothia (1975) and Li & Nussinov (1998). Richards' values are an adaptation of the values for individual atoms given by Bondi (1964). Chothia's values were derived from a very simple version of the calculations described here: an examination of the packing contacts found in the crystal structures of amino acids that were available when the work was carried out. Li & Nussinov (1998) derived their values from an examination of the distances between atomic groups in protein structures.

Comparison of the ProtOr radii with the radii determined by Chothia (1975) shows close agreement for aliphatic carbon (C4H3, C4H2, C4H1), aromatic carbon (C3H1) and nitrogen atoms (N3H2, N3H1). The ProtOr values for oxygen atom are somewhat higher than those used by Chothia (1975), while that for O2H1 is lower than that used by Richards (1974). The new ProtOr sulphur radius is lower than any of the previously determined values, but it is still within 0.08 to 0.17 Å of the others.

Here, we defined 13 different types of atoms. Li & Nussinov (1998) defined 24 types of atomic groups. They have, for example, three types of C4H2 groups that come from three classes of sidechains. This means that for six of the atom types defined here, they give more than one value (Table 2). Some of their values are close to those in the ProtOr set but most are larger. For five groups, the differences range between 0.10 and 0.20 Å. The main cause for these differences is probably that the positions of atoms in protein structures are less precisely determined than those in crystals of small organic molecules. Protein structures are refined using X-ray data with much lower resolution than that used to determine the structure of small molecules and, in almost all cases, only part of the contents of the crystal unit cell is included in the refinement.

A striking difference between the new data and previous results is the radius for C3H0. Our study found a value of 1.61 Å, which is 0.09 to 0.20 Å lower than previous values (Table 2). One reason is the influence of short interactions between carbonyl groups, in which the oxygen atom of one group makes a close approach to the carbon atom of the other. Recent work (Allen *et al*., 1998) has shown that these carbonyl ''stacking'' interactions are very common and energetically comparable to hydrogen bonds. The individual *d* values obtained for carbonyl···carbonyl contacts in this work are short (3.27 Å for C3H0···C3H0 and 3.01 Å for C3H0···O1H0), supporting this hypothesis. See also the work on C═O···C═O interactions in organic crystals by Bolton (1964) and in proteins by Maccallum *et al*. (1995a,b).

The determination of the volumes of atomic groups and residues

The Voronoi procedure for the calculation of volumes

The volumes of atomic groups were determined by the procedure originally developed by Voronoi (1908) and used by Bernal & Finney (1967) to study the structure of liquids. The method was first applied to proteins by Richards (1974). By constructing the minimally sized polyhedron (called a Voronoi polyhedron) around each atom, this procedure allocates the space within a structure, including cavities or defects, to its constituent atoms. As depicted in two dimensions by Figure 1, the faces of a Voronoi polyhedron are formed by planes perpendicular to lines between an atom and its neighbours, and the edges of a polyhedron result from the intersection of these planes. The volume inside a polyhedron is inversely proportional to the packing density of its central atom, i.e. a big volume is indicative of a loosely packed atom and a small volume a closely packed atom.

We used an implementation of the Voronoi procedure that is based on an original program by Richards (1974) with its subsequent modifications and extensions by Harpaz *et al*. (1994) and Gerstein *et al*. (1995). The full source code is available at http://bioinfo.mbb.yale.edu/geometry.

In the original Voronoi procedure the planes that form the polyhedron are placed at the midpoint of the lines between atoms and their neighbours and this allocates precisely all the space within a structure to its constituent atoms. This bisection method, however, cannot be used when different types of atoms have different radii. Placing the plane at the midpoint of the line between a small and large atom in direct contact plane will allocate part of the VDW envelope of the larger atom to the smaller atom. This problem can be overcome if the plane is placed not at the midpoint of the line, but at a position proportional to the radii of the two

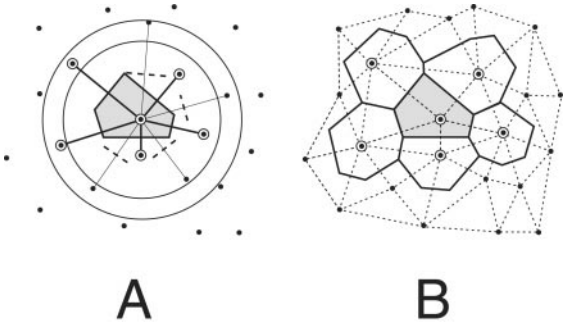

Figure 1. Two-dimensional representation of the Voronoi polyhedra and the Delaunay Tessellation calculations. (a) A polyhedron is built around the central atom. Points are the centres of atoms. Circled points are neighbours to the central atom. The calculation takes points within a distance cutoff (the outer circle) and creates faces between these atoms and the central atom. Neighbours share a face whereas atoms occluded by others do not (broken lines). The outer circle shows how a distance cutoff can overestimate neighbours. The inner circle shows how a distance cutoff can also underestimate neighbours. (b) Polyhedra are built around the same central polyhedra (shown in grey). The Delaunay Tessellation between atoms is shown by broken lines.

atoms joined by the line (Richards, 1974) This method of constructing the polyhedron introduces small errors in the calculations (Richards, 1985), since the planes no longer intersect at the same point. However, calculations show that these vertex errors are less than one part in 500 (Gerstein *et al*., 1995). Atoms in contact with each other were determined by the edges of the Delaunay triangulation (Delaunay, 1934).

In the calculations described here, we use the ProtOr VDW radii described above to position plane between different types of atoms. In principle, the position of planes between atoms that are hydrogen-bonded to each other should be determined using hydrogen bond radii rather than VDW radii. However, the ratios of the hydrogen bond radii for O1H0:O2H1:N3H1:N3H2:N4H3 is 1.00: 1.03 :1.16:1.20:1.12, which is close to the ratio of their VDW radii of 1.00:1.03:1.15:1.15:1.15. Thus the use of VDW radii to determine the position of planes between hydrogen-bonded polar atoms produces little or no error in their volumes. Also, computationally, it is much simpler.

Residue volumes were calculated by summing the volumes of their individual atoms. Standard deviations for these residue volumes were derived by taking the sum of the atom's deviations weighted by its contribution to the overall volume.

The protein structures used to determine the volume of atomic groups

To calculate the atomic volumes, we used protein structures that, at the time of this work, best represented protein residues in different environments: all the proteins have different folds and have been accurately determined at high resolution. The structures were selected using the SCOP database (Murzin *et al*., 1995). This database classifies the domains of all known protein structures in terms of their class, fold, superfamily, and family. We first extracted the best-determined structure for each of the 317 folds described in the version of SCOP available at the time this work was carried out. The experimental details of each structure were examined. Those that were found to have been determined at a resolution of 1.75 Å or better and to have *R* factors of 20 % or less were used in the calculations reported here. This procedure yielded 87 structures, and their Protein Data Bank identifiers (Abola *et al*., 1997; Bernstein *et al*., 1977) are listed in Table 4. In all, these structures contain 134,689 protein atomic groups, 1715 atomic groups from various types of ligands and cofactors, and 18,565 water molecules.

Different sets of buried atoms in proteins

The Voronoi procedure can only be used to calculate the volume of atoms that are surrounded by other atoms. This means that in proteins, we can calculate only the volume of atomic groups surrounded by other protein atoms, ligands, and water molecules whose positions in the interior or on the surface have been determined in the crystallographic analyses.

To answer the question of whether different regions in a protein have different packing densities, mean atomic volumes were calculated for selected sets of protein atoms. The sets of atoms for which volumes were determined are outlined in the following list.

B: This set contains protein atoms that are buried by other protein atoms and by ligands and/or cofactors. In selecting this set, the crystallographically determined water structure is ignored, i.e. the protein atoms used are those that have zero accessible surface area (Lee & Richards, 1971; Connolly, 1983) as calculated using just the atoms in the proteins, ligands, and cofactors.

BL: This set contains atoms that are buried as defined by the B set less those whose volumes are

Table 4. Structure set

Number	Identifier
87	1cbn, 1lkk, 2erl, 8rxn, 1bpi, 1ctj, 1igd, 1rge, 1amm, 1arb, 1cse, 1jbc, 2sn3, 1cus, 7rsa, 1rro, 1aac, 193l, 1utg, 5p21, 1hms, 1xyz, 256b, 2olb, 2phy, 3ebx, 3sdh, 2end, 1xso, 1cka, 1cyo, 1edm, 1ezm, 1isu, 1mla, 1poa, 1rie, 1whi, 2ctb, 2eng, 2ovo, 2cba, 3grs, 1lit, 1ra9, 1tca, 1csh, 1epn, 1mrj, 1phc, 1ptf, 1smd, 1vcc, 2dri, 2ilk, 2sil, 3pte, 4fgf, 2cpl, 1kap, 1lcp, 1php, 1snc, 1sri, 2wrp, 1krn, 2trx, 1ctf, 1fnb, 1gai, 1gof, 1knb, 1llp, 1mol, 1pdo, 1rop, 1tad, 1tfe, 1vhh, 1vsd, 2act, 1fkd, 1chd, 1kpt, 1thw, 2bbk, 3cla

affected by ligands and cofactors. The set was selected by removing from set B those atoms whose volumes are different when they are calculated in the presence and absence of ligands and cofactors. The L in the name of this set indicates this extra filtering of atoms.

BLW: This set contains atoms that are buried by other protein atoms less those whose volumes are affected by ligands and cofactors and by water molecules. The set was selected by removing from set BL those atoms whose volumes are different when they are calculated in the presence and absence of ligands, cofactors and water molecules. The set of volumes calculated from this set of atoms is given the label BLW.

BD: The atoms excluded from this set are (1) all those that have surface accessible to the solvent (as in set B) and (2) all those in contact to these surface atoms. Thus, both surface atoms and those that form the first layer below the surface are removed from the calculation to leave only those that are deeply buried. Therefore, the volumes produced by this set of atoms are named BD, where the D indicates that the resulting set of atoms are buried deep in the protein.

In our calculations, we used one extra step that affected the proteins volumes but did not affect the number of protein atoms in the set. For the B and BL sets, we calculated volumes with and without the water molecules whose position had been determined in the crystallographic analysis. A "+" sign added to a set's name indicates that we included the waters, and a "−" indicates that we did not. The reasons for carrying out these two calculations is described below in the next section. (Because of how the BLW and BL sets are defined, there are no differences between the volumes of the BLW+ and BLW− or BD+ and BD − .)

In the order given above, atoms in each set represent a progressively more deeply buried portion of the protein. This also means, of course, that we are selecting smaller and smaller numbers of atoms. The actual number of atoms in each set is shown in Table 5. The largest set of atoms contains more than three times as many atoms as the smallest.

Set B consists of protein atoms that are buried within the protein by other protein atoms and by ligands. Comparing the number of atoms in this set with the total number in the structures shows that the proportion of the atoms that have some access to the solvent is 100 % − 46 % = 54 %.

Table 5. Protein atoms in each set

Set	Number	%
Total protein atoms	134,689	100
B	61,786	46
BL	59,368	44
BLW	43,102	32
BD	19,510	15

Set BL contains atoms that are inaccessible to the solvent and whose volumes are not affected by ligand atoms. The proportion of atoms whose volumes are affected by ligands is small in the structures used here, 46 − 44 % = 2 %.

Set BLW contains atoms that are inaccessible to the solvent and whose volumes are not affected by ligand atoms or by the water molecules detected in the crystallographic analysis. The proportion of atoms whose volumes are affected by water molecules is 44 % − 32 % = 12 %. Given that this 12 % of atoms is inaccessible to solvent, it is perhaps surprising that they have volumes affected by the water structure. How this occurs is discussed in the section below on the role of water molecules in packing density of protein interiors.

Set BD excludes both surface atoms and those that form the first layer below the surface. This means that it contains only those that are deeply buried in the protein. In the structures considered here the number of such atoms is small: about one-seventh (14 %) of the total.

The volumes of atomic groups and residues in proteins

As described above, we carried out six different calculations for atomic volumes: on the BD and BLW set of atoms and on the BL and B sets with water molecules (BL+ and B+) and without water molecules (BL− and B−). The mean volume of atomic groups and residues produced by these six calculations are listed in Table 6. Data for 21 types of residues are given because the oxidised and reduced forms of cysteine, Cys and Cyh, are treated separately. In all, these 21 residues have 173 atomic groups.

Full details of the six volume calculations, i.e. the number of atoms used to compute each mean volume, the standard deviations of the mean atomic and residue volumes, and the range of the individual atomic volumes, are available at the web site http://bioinfo.mbb.yale.edu/geometry.

Standard deviations and the number of counts

In each set of calculations, the standard deviations of the mean residue volumes are between 2.4 and 4.4 %, with the exception of the following small residues: Gly where the range is 4.3 to 4.8 %, Cyh where it is 4.4 to 6.0 %, and Ser where it is 3.9 to 4.8 %.

The large majority of the mean volumes for atomic groups have standard deviations in the range of 6 to 11 %. Larger values are found for certain polar groups and a few of the adjacent carbon groups. These have standard deviations in the range of 12 to 17 %. There are 17 such atomic groups in set B, 13 in set BL, five in set BLW, and six in set BD.

For aliphatic and aromatic residues, the number of each of their atomic groups is high in all six sets of atoms: up about a 1000 for some groups in set

B, and 80 to 200 for most groups in set BD. For polar and charged residues, the situation is more complicated. The number of examples of their main-chain and aliphatic atomic groups is large in sets B, BL, and BLW. In most cases, one to several hundred. However, in set BD, it is usually between 50 and 100. The number of polar side-chain groups tends to be small and drops sharply on going from set B to set BD. This is especially the case for Lys N^{ζ}, which drops from 62 to six; Cyh S^{γ} which drops from 84 to 25, and Arg $N^{\eta 1}/N^{\eta 2}$ which drops from 187/145 to 28/21.

Differences in the residue and atomic volumes produced by ligand interactions

The B− and BL− sets (and the B+ and BL+ sets) differ in that the latter does not contain atomic groups whose volumes are affected by ligand interactions. For 19 residues the volumes given by the two sets are the same to within 0.4 $Å^3$ and, in most cases, within 0.2 $Å^3$. The exceptions are Cyh and His. In these residues the differences are mainly due to direct ligand interactions of the S^{γ} and $N^{\varepsilon 2}$ atoms. The unliganded form of these atoms in the absence of water have mean volumes of 37.0 and 16.4 $Å^3$, respectively, and the liganded atoms have mean volumes of 26.9 and 12.4 $Å^3$, respectively.

Residue and atomic volumes in the different sets

The six calculations produced six values for the volume of each residue and atomic group (Table 6). To a first approximation, the different values for a given residue are very similar. The values for aliphatic residues differ by no more than 0.5 to 1.6 %, and those for aromatic residues by no more than 0.5 to 1.1 % (if His is excluded). Polar residues (together with His) and charged residues display differences that are a little larger: 1.3 to 2.5 % in the first group and 2.0 to 3.5 % in the second.

Inspection of the differences in volumes for individual residues shows that though they are small, they do tend to be systematic:

$$BD \approx BLW \approx BL+ \approx B+ < BL- \approx B-$$

The volumes given for atomic groups and residues from the BD, BLW, BL+ and B+ calculations differ in most cases by less than 1 %. Only the volumes for Asn, Gln, Asp and Lys given by the BD set of atoms differ from those given by the BLW, BL+ and B+ sets by larger amounts: 1.9 to 2.7 %. However, the BD set possesses very few of these residues' side-chain atoms and these differences can not be considered significant.

The role of water in the packing density of protein interiors

The two calculations that give relatively high volumes (B− and BL−) use all, or almost all, buried atoms and do not include water in the volume calculation. If the crystallographic water molecules are included in the calculations, these sets give smaller volumes (B+ and BL+): the volumes for aliphatic and aromatic residues are 1 % smaller, on average, and those for polar and charged residues are 2 % smaller, on average. The volumes given by the B+ and BL+ calculations are mostly the same as those given by the two sets of deeply buried atomic groups, BD and BLW (see Table 6). Note that the crystallographic water structure plays no role in the BD and BLW calculations.

In the structures used here, the number of protein atoms that are inaccessible to solvent but have their volumes affected by the solvent is 16,266, i.e. a quarter of all the inaccessible protein atoms (see Table 5). The packing density of these atoms calculated in the presence of water is 3.4 % smaller than when it is calculated in absence of water

As mentioned above, it may seem contradictory that atoms which are not accessible to water should have their volumes affected by water. Water molecules in buried cavities account for a small fraction of this effect, but it is mostly due to the crystallographic water structure in cavities and grooves on the protein surface (Kuhn *et al.*, 1992; Gerstein & Chothia, 1996). Inspection of the buried atoms that are affected shows that, in many cases, they sit in cavities that contain water. These atoms do not make direct contact, but are second-nearest neighbours. In Figure 2, we provide a simple example of how measurements of their volume are affected by whether or not solvent is included in the calculation.

Figure 2 shows two surface atoms, a water molecule and an atom that is not directly in contact with water but does have it as a second-nearest neighbour. In the absence of water (Figure 2(a)), the Voronoi construction around the inaccessible atom produces a vertex ''pointing'' out toward solvent. In contrast, the presence of the water molecule, this vertex is ''clipped'' by a new polyhedra plane perpendicular to the ''long'' vector between the inaccessible atom and the water molecule (Figure 2(b)). This reduces the volume of the polyhedron around the accessible atom. (One can describe this situation somewhat more succintly in terms of the Delaunay tessellation, a geometric construction closely related to Voronoi polyhedra, as the introduction of the water molecule creating a new edge and fundementally changing the neighbours of each atom.)

This effect of water molecules on the volume calculation is illustrated by histograms of the volumes of four representative atom types displayed in Figure 3. We show the results from the BL− and BL+ calculations. Figure 3(a) shows the C3H1 atom-type distribution, for which there is the smallest difference in standard deviation between carbon atom types. The curves are essentially the same shape. In contrast, Figure 3(b) shows the

Table 6. The mean volume of the atomic groups and residues in different sets of protein atoms

(a) Aliphatic and Aromatic Residues

Res	Atm	Mean volume of atoms in each set						Res	Atm	Mean volume of atoms in each set					
		BD	BLW	BL+	B+	BL-	B-			BD	BLW	BL+	B+	BL-	B-
Gly	N	14.5	14.5	14.5	14.5	14.8	14.9	His	N	13.6	13.5	13.5	13.5	13.7	13.7
	Cα	23.3	23.5	23.5	23.4	23.8	23.7		Cα	13.3	13.2	13.3	13.3	13.5	13.5
	C	9.5	9.5	9.7	9.7	9.8	9.8		C	8.8	8.7	8.8	8.8	8.9	8.9
	O	16.2	16.2	16.2	16.1	16.5	16.4		O	16.2	15.9	15.9	15.8	16.1	16.1
	total	63.4	63.7	63.8	63.6	64.9	64.8		Cβ	23.2	23.5	23.4	23.3	23.8	23.7
Ala	N	13.8	13.8	13.9	13.9	14.0	14.0		Cγ	9.9	9.8	9.9	9.9	9.9	10.0
	Cα	14.0	14.0	14.0	14.0	14.1	14.1		Nδ1	15.1	15.4	15.5	15.2	15.8	15.5
	C	8.8	8.8	8.9	8.9	8.9	8.9		Cδ2	21.3	20.9	20.9	20.9	21.1	21.1
	O	16.0	16.0	16.0	16.0	16.2	16.3		Cε1	20.8	20.1	20.5	20.3	20.8	20.6
	Cβ	36.9	36.7	36.6	36.5	36.9	36.8		Nε2	14.9	15.6	15.8	14.8	16.4	15.3
	total	89.0	89.3	89.3	89.2	90.0	90.0		total	157.1	156.5	157.5	155.8	160.0	158.3
Val	N	13.6	13.5	13.6	13.6	13.7	13.7	Phe	N	13.6	13.5	13.5	13.5	13.7	13.7
	Cα	13.1	13.1	13.1	13.1	13.1	13.1		Cα	13.4	13.3	13.4	13.4	13.4	13.4
	C	8.5	8.5	8.5	8.5	8.6	8.6		C	8.6	8.6	8.7	8.7	8.8	8.8
	O	16.0	16.0	16.0	16.0	16.2	16.2		O	16.0	16.0	16.0	16.0	16.1	16.2
	Cβ	14.5	14.5	14.5	14.5	14.6	14.6		Cβ	23.6	23.7	23.6	23.6	23.8	23.8
	Cγ1	36.1	36.4	36.3	36.3	36.6	36.6		Cγ	9.7	9.6	9.7	9.7	9.7	9.7
	Cγ2	36.0	36.4	36.2	36.1	36.3	36.3		Cδ1	20.3	20.4	20.3	20.3	20.4	20.4
	total	137.7	138.3	138.2	138.1	139.0	139.1		Cδ2	21.0	21.0	20.9	20.9	21.1	21.1
Leu	N	13.7	13.5	13.5	13.5	16.7	13.7		Cε1	21.4	21.5	21.5	21.5	21.6	21.6
	Cα	13.1	13.1	13.1	13.1	13.1	13.1		Cε2	21.5	21.7	21.6	21.6	21.7	21.7
	C	8.7	8.7	8.8	8.8	8.8	8.8		Cζ	21.6	21.6	21.6	21.6	21.7	21.7
	O	16.0	16.0	16.0	16.0	16.2	16.2		total	190.9	190.9	190.8	190.8	191.9	191.9
	Cβ	22.8	22.8	22.8	22.8	23.0	23.0	Tyr	N	13.5	13.4	13.5	13.5	13.7	13.7
	Cγ	14.6	14.7	14.7	14.7	14.8	14.8		Cα	13.3	13.2	13.2	13.3	13.3	13.3
	Cδ1	37.4	37.4	37.2	37.3	37.4	37.4		C	8.8	8.6	8.7	8.7	8.8	8.8
	Cδ2	36.8	37.1	37.0	37.0	37.2	37.1		O	16.0	15.9	15.9	15.9	16.1	16.1
	total	163.0	163.1	163.1	163.1	164.0	164.1		Cβ	23.4	23.5	23.4	23.4	23.7	23.7
Ile	N	13.6	13.4	13.5	13.5	13.6	13.6		Cγ	9.6	9.6	9.7	9.7	9.7	9.7
	Cα	12.9	13.0	12.9	12.9	13.0	13.0		Cδ1	20.0	20.1	20.1	20.1	20.2	20.2
	C	8.5	8.4	8.4	8.4	8.5	8.5		Cδ2	20.8	20.7	20.6	20.6	20.7	20.7
	O	16.1	15.9	15.9	15.9	16.2	16.1		Cε1	20.4	20.6	20.5	20.5	20.7	20.7
	Cβ	14.2	14.2	14.1	14.1	14.2	14.2		Cε2	20.3	20.6	20.6	20.6	20.8	20.8
	Cg1	24.1	24.1	24.1	24.1	24.2	24.2		Cζ	9.8	9.8	9.9	9.9	9.9	9.9
	Cγ2	35.7	36.0	35.8	35.8	36.0	36.0		OH	18.8	18.5	18.5	18.4	19.4	19.2
	Cδ2	38.5	38.2	38.2	38.3	38.3	38.4		total	194.7	194.6	194.6	194.4	197.0	196.8
	total	163.6	163.2	163.0	163.1	163.9	164.0	Trp	N	13.7	13.5	13.6	13.6	13.8	13.8
Met	N	13.5	13.3	13.4	13.4	13.6	13.6		Cα	13.2	13.3	13.3	13.3	13.4	13.4
	Cα	13.2	13.2	13.2	13.2	13.2	13.2		C	8.7	8.6	8.7	8.7	8.7	8.7
	C	8.8	8.7	8.8	8.8	8.8	8.8		O	16.0	15.8	15.8	15.8	16.0	16.0
	O	15.9	15.9	16.0	16.0	16.3	16.3		Cβ	24.0	24.1	23.8	23.7	24.1	24.0
	Cβ	23.5	23.5	23.4	23.4	23.7	23.6		Cγ	9.8	9.9	9.9	9.9	10.0	10.0
	Cγ	24.0	23.7	23.8	23.9	24.0	24.0		Cδ1	20.5	20.5	20.6	20.5	20.9	20.9
	Sδ	29.6	30.3	30.2	30.0	30.3	30.1		Nε1	16.7	16.7	16.7	16.7	17.0	17.0
	Cε	36.8	36.8	37.0	36.9	37.3	37.2		Cδ2	10.1	10.0	10.1	10.1	10.1	10.1
	total	165.3	165.4	165.8	165.5	167.0	166.8		Cε2	9.7	9.8	9.8	9.8	9.9	9.9
Pro	N	8.5	8.6	8.7	8.6	8.8	8.8		Cε3	20.3	20.4	20.4	20.4	20.5	20.4
	Cα	13.9	13.8	13.8	13.8	14.0	14.0		Cζ3	21.4	21.5	21.4	21.4	21.5	21.5
	C	8.7	8.8	8.8	8.8	8.8	8.8		Cζ2	20.7	20.7	20.9	21.8	21.2	21.0
	O	16.3	15.8	15.9	15.9	16.2	16.2		Cη2	21.1	21.3	21.2	21.2	21.3	21.3
	Cβ	25.3	25.3	25.3	25.3	25.6	25.6		total	226.0	225.9	226.4	226.1	228.2	228.0
	Cγ	25.8	25.2	25.5	25.6	26.7	25.8								
	Cδ	23.6	23.6	23.4	23.4	23.8	23.8								
	total	122.1	121.1	121.6	121.4	122.9	123.0								

Table 6. *Continued.*

(b) Polar and Charged Residues

Res	Atm	Mean volume of atoms in each set BD	BLW	BL+	B+	BL-	B-
Cyh	N	14.0	13.8	13.9	13.9	14.0	14.0
	Cα	13.4	13.7	13.6	13.5	13.7	13.7
	C	8.8	8.7	8.8	8.8	8.9	8.9
	O	16.3	16.4	16.4	16.3	16.5	16.5
	Cβ	23.1	23.6	23.5	23.4	23.7	23.5
	Sγ	35.3	36.8	36.7	34.4	37.0	34.6
total		110.9	113.0	112.8	110.3	113.7	111.2
Cys	N	13.8	13.6	13.6	13.7	13.7	13.8
	Cα	13.2	13.0	13.1	13.1	13.1	13.2
	C	8.7	8.7	8.7	8.8	8.8	8.8
	O	16.3	16.0	16.1	16.0	16.2	16.2
	Cβ	23.4	23.2	23.4	23.4	23.6	23.6
	Sγ	27.7	27.6	27.5	27.5	27.7	27.7
total		103.0	102.0	102.5	102.0	103.3	103.2
Ser	N	13.9	13.8	13.8	13.8	14.1	14.1
	Cα	13.4	13.3	13.4	13.4	13.5	13.5
	C	8.8	8.8	8.9	8.9	8.9	8.9
	O	16.1	15.9	15.9	15.8	16.2	16.2
	Cβ	23.7	23.2	23.6	23.6	23.9	23.9
	Oγ	18.9	18.1	18.0	17.9	18.8	18.7
total		94.0	93.1	94.2	93.4	95.4	95.3
Thr	N	13.6	13.5	13.5	13.5	13.9	13.8
	Cα	13.2	12.9	13.0	13.0	13.1	13.1
	C	8.7	8.6	8.7	8.7	8.7	8.7
	O	16.0	15.8	15.8	15.8	16.0	16.0
	Cβ	14.8	14.8	14.7	14.7	14.9	14.8
	Oγ1	17.9	17.9	17.6	17.5	18.3	18.1
	Cγ2	35.8	36.2	36.3	36.2	36.6	36.5
total		120.1	119.6	119.6	119.4	121.5	121.2
Asn	N	13.9	13.5	13.5	13.5	13.7	13.7
	Cα	13.1	13.0	13.1	13.0	13.2	13.2
	C	8.9	8.8	8.9	8.9	8.9	8.9
	O	16.3	16.0	15.9	15.9	16.2	16.2
	Cβ	23.4	22.6	22.7	22.8	23.1	23.1
	Cγ	9.5	9.4	9.5	9.5	9.6	9.6
	Oδ1	16.8	16.1	16.2	16.2	16.7	16.7
	Nδ2	23.3	22.6	22.5	22.3	23.3	23.1
total		125.0	121.9	122.4	122.1	124.7	124.5
Gln	N	13.7	13.3	13.4	13.4	13.6	13.6
	Cα	13.3	13.2	13.2	13.2	13.3	13.3
	C	8.7	8.7	8.7	8.7	8.8	8.8
	O	15.9	15.8	15.8	15.8	16.0	16.0
	Cβ	23.2	22.9	23.1	23.1	23.3	23.3
	Cγ	23.6	23.1	23.2	23.4	23.4	23.5
	Cδ	9.7	9.4	9.6	9.6	9.7	9.7
	Oε1	17.1	16.7	16.6	16.5	17.2	17.1
	Nε2	23.6	22.5	23.3	23.1	24.1	24.2
total		148.7	145.7	146.9	146.7	149.4	149.4
Asp	N	13.9	13.6	13.7	13.7	13.9	13.9
	Cα	13.2	13.2	13.3	13.2	13.4	13.4
	C	8.7	8.7	8.8	8.8	8.8	8.8
	O	16.3	15.8	15.8	15.8	16.1	16.1
	Cβ	23.0	22.5	23.0	23.2	23.4	23.5
	Cγ	9.2	9.2	9.3	9.4	9.4	9.4
	Oδ1	15.3	15.1	15.1	14.9	15.6	15.5
	Oδ2	15.9	15.2	15.6	15.3	16.7	16.3
total		115.5	113.3	114.4	114.2	117.3	116.9
Glu	N	13.7	13.4	13.5	13.5	13.6	13.6
	Cα	13.4	13.3	13.3	13.3	13.4	13.4
	C	8.6	18.6	8.6	8.6	8.7	8.7
	O	15.7	15.7	15.8	15.8	16.0	16.0
	Cβ	23.3	23.4	23.2	23.2	23.5	23.5
	Cγ	23.0	23.1	23.3	23.3	23.5	23.5
	Cδ	9.4	9.4	9.4	9.4	9.5	9.5
	Oε1	15.4	15.5	15.5	15.5	16.0	16.0
	Oε2	15.8	16,3	16.2	16.0	17.9	17.4
total		138.3	138.7	138.8	138.6	142.2	141.6
Lys	N	13.3	13.3	13.4	13.4	13.7	13.7
	Cα	13.4	13.2	13.2	13.2	13.3	13.3
	C	8.7	8.6	8.7	8.7	8.8	8.8
	O	16.2	15.9	15.8	15.8	16.1	16.0
	Cβ	23.0	22.6	22.6	22.7	22.8	22.9
	Cγ	23.7	23.3	22.8	22.9	23.1	23.1
	Cδ	24.1	22.5	23.4	23.5	23.7	23.8
	Cε	23.3	23.2	23.7	23.5	24.0	23.9
	Nζ	21.7	21.6	21.4	20.6	21.9	21.2
total		168.5	164.1	165.1	164.4	167.3	166.9
Arg	N	13.6	13.4	13.5	13.5	13.6	13.7
	Cα	13.4	13.3	13.3	13.3	13.4	13.4
	C	8.8	8.7	8.8	8.8	8.8	8.8
	O	16.0	15.9	15.9	15.9	16.1	16.1
	Cβ	22.9	22.6	22.8	22.8	23.1	23.1
	Cγ	23.2	23.3	23.3	23.3	23.6	23.6
	Cδ	23.3	23.1	22.8	22.9	23.3	23.3
	Nε	15.2	14.9	15.0	15.0	15.5	15.6
	Cζ	9.6	9.5	9.7	9.7	9.7	9.7
	Nη1	22.3	21.9	22.1	22.0	22.8	22.8
	Nη2	23.2	23.7	23.1	23.1	23.9	23.8
total		191.5	190.2	190.3	190.3	194.0	193.9

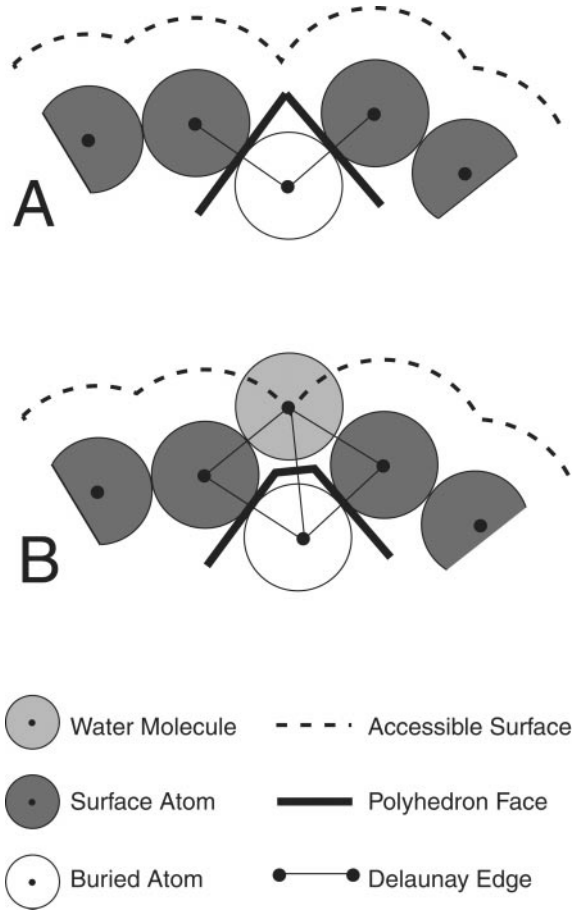

Figure 2. Water clipping. This Figure illustrates how the volume of a buried protein atom can be affected by the presence of nearby water molecules. In the diagram the buried protein atom is indicated by the white, open circle. It is considered to be ''buried'' according to the Lee & Richards (1971) definition: an atom is buried if it cannot make direct contact with a water molecule. (a) A polyhedron around this buried atom. (b) A polyhedron around the same atom with the addition of a nearby water molecule (light gray). Notice how the water molecule ''clips'' the polyhedron, reducing its volume. Fundamentally, this situation arises because of the difference between the way the protein ''surface'' is defined by the Voronoi construction and by the Lee & Richards probe sphere. The Voronoi construction defines the protein surface in terms of the sharing of a polyhedron ''face'' with a solvent molecule. This is equivalent to whether a protein atom has a Delaunay edge connecting it with a solvent molecule. (These Delaunay edges are indicated in the diagram by the thin continuous lines.) Lee & Richards (1971), in contrast, define the surface in terms of contact with an imaginary probe sphere.

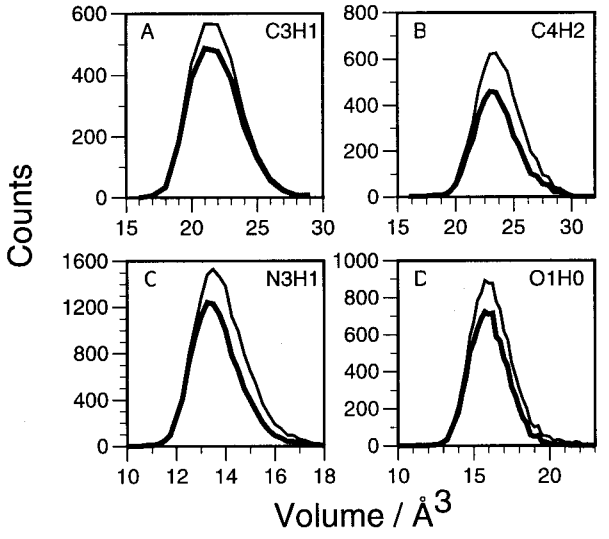

Figure 3. Atom type histograms. Distribution of volumes according to atomic group are shown for two types of protein sets: BL− and BL+ (see the text for an explanation of protein sets). To facilitate comparisons, the distributions were not normalised and the atoms were taken from the following explained subsets of atoms. (a) Aromatic carbon atoms with one hydrogen atom from Phe and Trp. (b) All aliphatic carbon atoms with two hydrogen atoms except the following: Cys C^{β}, Cyh C^{β}, Ile $C^{\gamma 1}$, Lys C^{ε}, Met C^{γ}, Pro C^{β}, and Pro C^{γ}. (c) Amide main-chain nitrogen atoms with one hydrogen atoms except the one from Pro. (d) All carboxyl or carbonyl oxygen atoms.

C4H2 atom-type distribution, for which there is the greatest difference in standard deviation between carbon atom types. Besides height, the curve from the BL− set shows a pronounced skewing towards larger volumes. Including these larger values causes the increase in the atom type's average volume and standard deviation. In Figure 3(b) and (c) we repeat the comparison for two polar atoms. Both show the same effects when the crystallographically determined water molecules are included in the calculation: atomic group volumes decrease.

General uniformity of the packing density in the interior of proteins

The results presented above demonstrate that the average packing density in the interior of proteins is high and that there are no systematic differences in the density of different regions of the interior. In the deep interior there are few or no cavities. Towards the surface, cavities and grooves do occur, but water effectively fills the spaces and produces a density in these regions very similar to that in the deep interior. Our results should not be taken to imply that the absolute density is the same at all points in the interior. Specifically, regions that contain polar atoms joined by short hydrogen bonds have a higher absolute density than regions that contain aliphatic atoms interacting through the longer VDW contacts (Kuntz, 1972). Also, significant cavities are found occasionally in protein interiors. However, these local variations are not systematic and they do not affect the general conclusion that the efficiency of the different types of packing, whether through hydrogen

bonds or VDW interactions, is very similar throughout the protein interior.

Comparison with other published values for residue and atomic volumes

Values for the mean volumes of residues buried in proteins have been published by Harpaz *et al*. (1994) and by Pontius *et al*. (1996) who also list values for the mean volumes of atomic groups. These two sets of residue volumes are listed in Table 7. None of the procedures for calculating volumes used in this study correspond exactly to those used in the two previous calculations, but the BL− calculation, which uses all buried atoms not in contact with ligands and with no water molecules involved in the calculation, comes the closest. In Table 7, we list the BL− residue volumes and the percentage differences that they have with the two previous sets.

In comparison with our calculations, Harpaz *et al*. (1994) used the Chothia (1975) atomic radii (which for certain atoms are different to those used here; see Table 2); selected a different set of protein structures; and counted only the volumes of residues that have all atoms buried. In spite of these differences, the mean residue volumes are close to the BL− set for most residues: ten residues differ by 0.0 to 0.5 %, and five by 0.5 to 1.0 % (Table 7).

Larger differences are found for Ser (−1.3 %), Thr (−1.2 %), Gly (−1.7 %), Trp (+1.5 %), Lys (+1.6 %) and Asn (+2.2 %). Inspection of their atomic volumes shows that these larger differences mainly arise from the use of different atomic radii in the two calculations. Ser, Thr, and Gly have a high proportion of oxygen atoms and the differences are produced mainly by the different radii used for the O1H0 and O2H1 groups: 1.42 Å and 1.46 Å, respectively, in this study and 1.40 Å for both groups by Harpaz *et al*. (1994). Similarly, the differences for Trp are produced mainly by the different radius used for the four C3H0 groups in this residue: 1.61 Å in this work and 1.76 Å in the work by Harpaz *et al*. (1994). The differences for Lys and Asn are not related simply to different atomic radii differences and probably reflect the small number of residues used to determine the Harpaz *et al*. (1994) values, 6 and 41, respectively.

The residue volumes calculated by Pontius *et al*. (1996), compared to those calculated here, have large differences: −3 % to +15 % (see Table 7). The main reason for this is that the method used to allocate space to the Voronoi polyhedra is not the same in the two calculations. As shown in Figure 1, the faces of a Voronoi polyhedron are formed by planes perpendicular to vectors between an atom. Its neighbours and the edges of a polyhedron result from the intersection of these planes. In the calculations described here, planes between different types of atoms are placed at positions proportional to their VDW radii. Pontius *et al*. (1996) place the planes at the midpoint of the vectors between atoms regardless of their chemical type. This was because they wished to obtain a set of atomic and residue volumes that are independent of any particular set of van der Waals radii, an important consideration for accomodating exotic ligands. As pointed out earlier, placing the plane at the midpoint of the lines between atoms tends to transfer parts of the VDW envelope of large atoms to those small atoms with which they are in direct contact. This means that it produces larger

Table 7. Comparison of the residue volumes with previous calculations

Residue	Residue volumes (Å^3)			Volume differences (%)	
	Harpaz[a] (A)	Pontius[b] (B)	BL− (C)	100(A-C)/C	100(B-C)/C
Gly	63.8	67.5	64.9	−1.7	+4.0
Ala	90.1	91.5	90.0	+0.1	+1.7
Val	139.1	138.4	139.0	+0.1	−0.4
Leu	164.6	163.4	164.0	+0.4	−0.4
Ile	164.9	162.6	163.9	+0.6	−0.8
Met	167.7	165.9	167.0	+0.4	−0.7
Pro	123.1	123.4	122.9	+0.2	+0.4
His	159.3	162.3	160.0	−0.4	+1.4
Phe	193.5	198.8	191.9	+0.8	+3.6
Tyr	197.1	209.8	197.0	+0.1	+6.5
Trp	231.7	237.2	228.2	+1.5	+3.9
Cyh	113.2	114.4	113.7	−0.4	+0.6
Cys	103.5	102.4	103.3	+0.2	−0.9
Ser	94.2	102.0	95.4	−1.3	+6.9
Thr	120.0	126.0	121.5	−1.2	+3.7
Asn	127.5	138.3	124.7	+2.2	+10.9
Gln	149.4	156.4	149.4	0.0	+4.7
Asp	117.1	135.2	117.3	−0.2	+15.3
Glu	140.8	154.6	142.2	−1.0	+8.7
Lys	170.0	162.5	167.3	+1.6	−2.9
Arg	192.8	196.1	194.0	−0.6	+1.1

[a] Harpaz *et al*. (1994).
[b] Pontius *et al*. (1996).

volumes for groups that in our calculations are small and *vice versa*. For example, the C4H3 and main-chain O1H0 groups have volumes of close to 37 Å^3 and 16 Å^3 here and close to 34 Å^3 and 22 Å^3 in the calculations by Pontius *et al*. (1996). The greatest net effects are on residues that have a high proportion of oxygen atoms (Table 7). Thus, the volume given by Pontius *et al*. (1996) for the residue Asp, 135 Å^3, is 11% larger than that given here. It is also a little larger than the volume that the amino acid occupies in its crystal, 133.1 Å^3 (Chothia, 1975).

Calculation of the partial specific volumes of proteins using the mean volumes of buried residues

Harpaz *et al*. (1994) found that the volumes proteins occupy in solution can be calculated quite accurately using their amino acid composition and the mean volumes of buried residues. The calculated values differed from the experimental ones by −0.5% on average. This rather surprising result, that the volume of a protein in solution can be calculated almost exactly by using for all residues the volumes occupied by buried residues, was given a semi-quantitative explanation in subsequent work by Gerstein & Chothia (1996). They showed that at the protein surface: (1) protein atoms occupy on average volumes that are larger than those in the interior by an amount that is roughly in proportion to their exposure to solvent; (2) water molecules occupy on average volumes that are smaller than those in the bulk solvent by an amount proportional to the extent of their protein contacts; and (3) these volume changes tend to cancel each other.

Do the residue volumes described here support the conclusions of the previous work? Experimental determinations of the protein volumes in solution are measured as partial specific volumes: $\overline{v}$ their volumes divided by their molecular masses. To answer this question, we used the residue volumes determined here to calculate the $\overline{v}$ of proteins for which the calculation was carried out by Harpaz *et al*. (1994) and compared them to the experimental values. To calculate volumes, we also used what we believe to be the most appropriate values for buried residues in proteins: the BL+ set listed in Table 6.

The values for partial specific volumes of 12 proteins were calculated by a procedure similar to the one used by Harpaz *et al*. (1994): (1) For each residue in a sequence the volumes and molecular masses were summed. (2) Adjustments were made for the groups at the chain termini. For the terminal carboxyl, an average value for a carboxyl oxygen of 18.0 Å^3 was added. For the terminal amino group, the volume difference between an amino and amide groups of 1.3 Å^3 was subtracted (Table 6). (3) To compensate for the electroconstriction of the charged groups, 10 Å^3 was taken off the volume for each carboxyl oxygen atom and 18 Å^3 for each amino or guanidino group (Cohn & Edsall, 1943; Mishra & Ahluwalia, 1984). (4) To get the partial specific volume, we divided the total volume by the total mass and used the conversion factor of 0.6023 to change the units of Å^3/a.m.u. to the more conventionally used ml/g.

The calculated $\overline{v}$ values are given in Table 8.

Experimental values, taken from the compilation by Squire & Himmel (1979) and Gekko & Noguchi (1979), are also given in Table 8. The partial specific volumes calculated by Harpaz *et al*. (1994) are on average 0.5% smaller than the experimental values. The new residue volumes used here give calculated values that are between 1.2 and 3.9% smaller than the experimental values, and 2.5% smaller on average.

These results modify only slightly the conclusions by Harpaz *et al*. (1994) and Gerstein & Chothia (1996). They imply that the increase in the volume occupied by atoms on the protein surface (relative to what they occupy in the interior) is not quite balanced by the decrease in volume occupied by the water molecules on the protein surface (relative to that in the bulk solvent). The net effect is to give a protein a volume ~2.5% larger than it would have if all protein atoms did occupy the same volume as interior atoms.

Table 8. Comparison of partial specific volumes

Protein	Source	Id	Partial specfic volume (ml/g)	
			Calculated	Experiment[a]
Alchohol dehydrogenase	Equine	1lde	0.721	0.750
Carbonic anhydrase B	Human	2cab	0.711	0.729
Carboxypeptidase A	Bovine	2ctb	0.715	0.733
Chymotrypsinogen	Bovine	2cga	0.721	0.732
Concanavalin A	Jack bean	1scs	0.713	0.732
Elastase	Porcine	1lvy	0.719	0.730
Hemoglobin	Equine	1mhb	0.722	0.750
Lysozyme	Chicken	8lyz	0.699	0.712
Malate dehydrogenase	Porcine	1mld	0.727	0.742
Ribonuclease A	Bovine	1xps	0.693	0.703
Subtilisin	*B. Amyl.*	1sbt	0.722	0.731
Superoxide dismutase	Bovine	1sda	0.706	0.729

[a] Squire & Himmel (1979); Gekko & Noguchi (1979).

Conclusions

We have derived a new set of radii and volumes for atomic groups in proteins. We expect that our ProtOr radii and the volumes for atomic groups and residues, particularly the BL+ and BL− values, will be useful in many potential applications. During the course of this study, we have examined the differences in packing density in various parts of the interior of proteins. If the water structure is ignored, atoms deeply buried in the protein interior are packed a little more densely than those nearer the protein surface. The water structure that fills the cavities and grooves on the protein surface, produces a packing density for atoms near the surface that is the same as that of deeply buried atoms. Elsewhere it has been shown that water molecules play a similar role in protein-protein and protein-DNA recognition sites (Lo Conte *et al*., 1999; Nadassy *et al*., 1999).

Acknowledgements

J.T. thanks the National Institutes of Health (grant number GM41455). M.G. thanks the NSF for support (grant DBI-9723182). We thank Ronald Jansen, Hedi Hegyi, Vadim Alexandrov, Michael Levitt, and David Baker for suggestion, support, and help.

References

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol.* **277**, 556-571.

Allen, F. H., Davies, J. E., Gallo, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. (1991). The development of versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **31**, 187-204.

Allen, F. H., Baalham, C. A., Lommerse, J. P. M. & Raithby, P. R. (1998). Carbonyl-carbonyl interactions can be competitive with hydrogen bonds. *Acta Crystallog. sect. B,* **54**, 320-329.

Bernal, J. D. & Finney, J. L. (1967). Random close-packed hard-sphere model II. Geometry of random packing of hard spheres. *Discus. Faraday Soc.* **43**, 62-69.

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

Bolton, W. (1964). Intermolecular carbonyl carbon-oxygen interactions in organic crystals. *Nature,* **201**, 987-989.

Bondi, A. (1964). VDW volumes and radii. *J. Phys. Chem.* **68**, 441-451.

Chothia, C. (1975). Structural invariants in protein folding. *Nature,* **254**, 304-308.

Cohn, E. J. & Edsall, J. T. (1943). *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, pp. 370-381, Reinhold, New York.

Connolly, M. (1983). Analytical molecular surface calculation. *J. Appl. Crystallog.* **16**, 548-558.

Delaunay, B. (1934). Sur la sphére vide. *Bull. Acad. Sci USSR (VII), Classe Sci. Mat. Nat.* 783-800.

Gekko, K. & Noguchi, H. (1979). Compressibility of globular proteins in water at 25°C. *J. Phys. Chem.* **83**, 2706-2714.

Gerstein, M. & Chothia, C. (1996). Packing at the protein-water interface. *Proc. Natl Acad. Sci. USA,* **93**, 10167-10172.

Gerstein, M., Tsai, J. & Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.* **249**, 955-966.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). Volume changes on protein folding. *Structure,* **2**, 641-649.

Kocher, J. P., Prevost, M., Wodak, S. J. & Lee, B. (1996). Properties of the protein matrix revealed by the free energy of cavity formation. *Structure,* **4**, 1517-1529.

Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzhoff, E. D. & Tainer, J. A. (1992). The interdependence of protein surface-topography and bound water-molecules revealed by surface accessibility and fractal density measures. *J. Mol. Biol.* **228**, 12-22.

Kuntz, I. D. (1972). Tertiary structure in carboxypeptidase. *J. Am. Chem. Soc.* **94**, 8568-8572.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.

Li, A. J. & Nussinov, R. (1998). A set of VDW and columbic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins: Struct. Funct. Genet.* **32**, 111-127.

Liang, J., Edelsbrunner, E. & Wodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications of ligand design. *Protein Sci.* **7**, 1884-1887.

Lo, Conte L., Chothia, C. & Janin, J. (1999). Atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177-2198.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995a). Columbic attractions between partially charged main-chain atoms stabilise the right-handed twist found in most beta-strands. *J. Mol. Biol.* **248**, 374-384.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995b). Columbic interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of alpha-helices and antiparallel beta-sheet. A new method for analysing the forces between hydrogen bonding groups in proteins includes all the Columbic interactions. *J. Mol. Biol.* **248**, 361-373.

Mishra, A. K. & Ahluwalia, J. C. (1984). Apparent molal volumes of amino acids, N-acetylamino acids and peptides in aqueous solutions. *J. Phys. Chem.* **88**, 86-92.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 537-540.

Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry,* **38**, 1999-2017.

Peters, K. P., Fauck, J. & Frommel, C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256**, 201-213.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). Deviations from standard atomic volumes as a quality measure of protein crystal structures. *J. Mol. Biol.* **264**, 121-136.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.

Richards, F. M. (1979). Packing defects, cavities, volume fluctuations, and access to the interior of proteins. including some general comments on surface area and protein structure. *Carlsberg. Res. Commun.* **44**, 47-63.

Richards, F. M. (1985). Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol.* **115**, 440-464.

Rowland, R. S. & Taylor, R. (1996). Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from VDW radii. *J. Phys. Chem.* **100**, 7384-7391.

Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.* **221**, 327-346.

Squire, P. G. & Himmel, M. E. (1979). Hydrodynamics and protein hydration. *Arch. Biochem. Biophys.* **196**, 165-177.

Voronoi, G. F. (1908). Nouvelles applications des paramétres continus á la théorie de formes quadratiques. *J. Reine Agnew. Math.* **134**, 198-287.

Edited by J. M. Thornton

(Received 17 December 1998; received in revised form 13 April 1999; accepted 26 April 1999)