BIOINFORMATICS

Vol. 18 no. 7 2002
Pages 985–995

Calculations of protein volumes: sensitivity analysis and parameter database

Jerry Tsai[1,*] *and Mark Gerstein*[2]

[1]*Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843-2128, USA and* [2]*Department of Molecular Biophysics and Biochemistry, Yale University, Bass Center, 266 Whitney Avenue, P.O. Box 208114, New Haven, CT 06520-8114, USA*

Received on October 30, 2001; revised on May 23, 2001; October 14, 2001; accepted on December 11, 2001

ABSTRACT

Motivation: The precise sizes of protein atoms in terms of occupied packing volume are of great importance. We have previously presented standard volumes for protein residues based on calculations with Voronoi-like polyhedra. To understand the applicability and limitations of our set, we investigated, in detail, the sensitivity of the volume calculations to a number of factors: (i) the van der Waals radii set, (ii) the criteria for including buried atoms in the calculations or atom selection, (iii) the method of positioning the dividing plane in polyhedra construction, and (iv) the set of structures used in the averaging.

Results: We find that different radii sets have only moderate affects to the distribution and mean of volumes. Atom selection and dividing plane methods cause larger changes in protein atoms volumes. More significantly, we show how the variation in volumes appears to be clearly related to the quality of the structures analyzed, with higher quality structures giving consistently smaller average volumes with less variance.

Availability/Supplementary Information: Programs and associated data files are available from http://bioinfo.mbb.yale.edu/geometry and http://molmovdb.org. In particular, we make available an extensive database of many different sets of protein geometric parameters.

Contact: JerryTsai@TAMU.edu; Mark.Gerstein@Yale.edu

INTRODUCTION

Increasing numbers of protein structures are solved every year (Thorton, 1992) and deposited in the Protein Data Bank (PDB) (Abola *et al.*, 1997; Bernstein *et al.*, 1977). To understand the structural elements of protein packing, volumes and radii of protein atoms need to be calculated. Unfortunately, structures solved using x-ray diffraction do not usually resolve the hydrogens. As a result, protein volumes and radii are usually calculated based on an atom group, where a heavy atom and its associated hydrogens are unified into a single entity. We will be discussing atomic group volumes throughout this paper. To simplify the language somewhat, we will refer to them as atom volumes—though we ask the reader to keep in mind that many of our atoms are really atom groups.

Calculating volumes and radii for these atoms is not straightforward, since they occupy irregular packing volumes and cannot be treated as simple spheres. This complication has been overcome using a number of methods, and volumes for atom groups have been calculated (Bondi, 1964; Chothia, 1974; Finney, 1975; Harpaz *et al.*, 1994; Li and Nussinov, 1998; Liang *et al.*, 1998a; Richards, 1974). Table 1 displays a standard set of protein atom and residue volumes determined from an analysis of high-resolution protein structures (Tsai *et al.*, 1999). Radii and volume sets are commonly used to characterize a number of protein properties, such as: protein energies (Chothia, 1975), protein–protein interactions (Janin and Chothia, 1990), standard residue volumes (Harpaz *et al.*, 1994), internal core packing (Janin, 1979; Richards, 1985), packing at the water interface (Gerstein and Chothia, 1996; Gerstein *et al.*, 1995), protein cavities (Hubbard and Argos, 1995; Liang *et al.*, 1998a,b; Richards, 1979), the quality of crystal structures (Pontius *et al.*, 1996), analysis of amino acid compositions (Gerstein, 1998; Gerstein *et al.*, 1994), macromolecular motions (Gerstein and Krebs, 1998; Krebs and Gerstein, 2000) and even measurement of the fit between an enzyme and its substrate (David, 1988; Finney, 1978). Standard volumes and radii are also important in an indirect sense in the prediction of side-chain packing (Dunbrack, 1999; Koehl and Delarue, 1997; Lee and Levitt, 1997).

One volume determination method constructs polyhedra around atoms (Voronoi, 1908). Bernal and Finney (1967) initially applied this calculation to molecular systems, and Richards (1974) used it first with proteins. Here, we

*To whom correspondence should be addressed.

ProtOr Volumes and Parameters

Unified atoms			Residues	
Atom	Radii	Volume	aa	Volume
C3H0b	1.61	9.70	Gly	63.8
C3H0s	1.61	8.72	Ala	89.3
C3H1b	1.76	21.28	Val	138.2
C3H1s	1.76	20.44	Leu	163.1
			Ile	163.0
C4H1b	1.88	14.35	Met	165.8
C4H1s	1.88	13.17		
C4H2b	1.88	24.26	Pro	121.6
C4H2s	1.88	23.19	His	157.5
C4H3u	1.88	36.73	Phe	190.8
			Tyr	194.6
N3H0u	1.64	8.65	Trp	226.4
N3H1b	1.64	15.72		
N3H1s	1.64	13.62	Cyh	112.8
N3H2u	1.64	22.69	Cys	102.5
			Ser	94.2
N4H3u	1.64	21.41	Thr	119.6
			Asn	112.4
O1H0u	1.42	15.91	Gln	146.9
O2H1u	1.46	17.98	Asp	114.4
			Glu	138.8
S2H0u	1.77	29.17	Lys	165.1
S2H1u	1.77	36.75	Arg	190.3

Parameters used in Protor Volume Derivation

Typing schem e	Hybrid chemical and numerical typing with 18 basic types
Radii set	ProtOr radii, Tsai et al. (1999)
Plane-positioning method	Ratio
Atom selection criteria	BL+
Structure set	SCOP (87 structures)

Table 1.

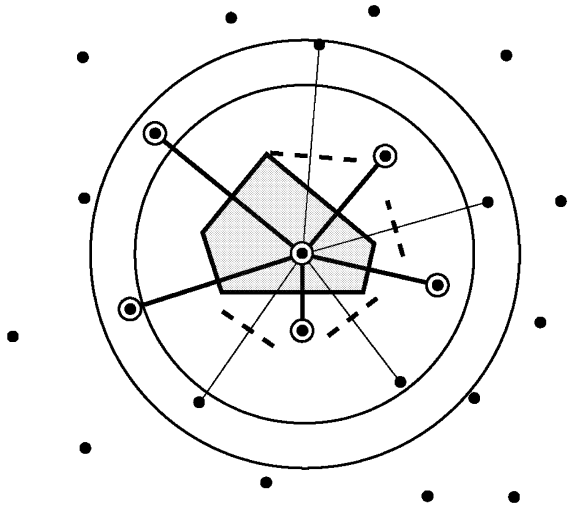

Fig. 1. Two-dimensional representation of calculating Voronoi polyhedra and the Delaunay tessellation. A polyhedron is built around the central atom. Points are the centers of atoms. Circled points are neighbors to the central atom. The calculation takes points within a distance cutoff (the outer circle). For each atom paired with the central atom, a face is created perpendicular to the line connecting the two atoms. These faces intersect to define a polyhedron's vertices. The faces and vertices form the boundary of the polyhedron (shown by the shaded area). Neighbors share a face (bold connecting lines and faces) whereas atoms occluded by others do not (light connecting lines and broken lines for faces). The outer circle shows how a distance cutoff can overestimate neighbors. The inner circle shows how a distance cutoff can also underestimate neighbors. The bold lines connecting the central atom to its neighbors are the Delaunay tessellation.

attempt to understand the sensitivity of the parameters that influence the calculation of polyhedra and extend our previous work (Gerstein and Chothia, 1996; Harpaz *et al.*, 1994; Tsai *et al.*, 1999). Figure 1 illustrates how a Voronoi polyhedron is built. The construction partitions space such that all points within a polyhedron are closer to its atom than any other. This partitioning provides a good estimate of an atom's true volume. The Delaunay triangulation is associated with the construction of polyhedra and is quite useful for unambiguously determining the contact neighbors of a given atom. Figure 1 also displays advantages of the triangulation over using a simple distance cutoff for determining neighbors. Because only neighbors share a polyhedron face or a Delaunay connection, there is no over- or underestimation of contacts in comparison to methods based on a cutoff radius.

Overview of our sensitivity analysis

Using Voronoi-like polyhedra, we previously determined standard sets of protein volumes in an analysis of high-resolution protein structures (Tsai *et al.*, 1999). These standards serve as valuable references in packing calculations. We call our standard reference volumes the ProtOr set. Our original ProtOr set included volumes for all 173 possible protein atoms (Tsai *et al.*, 1999). The 173 atoms are derived from considering each atom in 21 protein residues as distinct (167 atoms from the 20 regular amino acids and six more from an oxidized

Type	Chemistry	CH3 Aliphatic methyl	CH2 Aliphatic methyl	>CH Aliphatic CH	CH Aromatic CH	>C Trigonal aromatic	NH3+ Amino charged	NH2 Amino or amide	>NH Peptide, NH or N	OO Carbonyl Oxygen	HO Alcoholic hydroxyl	MS Carboxyl Oxygen	HS Sulfhydryl	Thioether or –S-S-	Notes
Bondi	1968	2.00	2.00	--		1.74	-	1.75	1.65	1.50	---			1.80	Values Assigned on bassis of observed packing in condensed phases (Bondi, 1968)
Lee & Richards	1971	1.80	1.80	1.70	1.80	1.80	1.80	1.80	1.52	1.80	1.80	1.80	1.80	-	Values adapted from Bondi (1964) and used in Lee and Richards (1971)
Shrake & Rupley	1973	2.00	2.00	2.00	1.85	*	1.50	1.50	1.40	1.40	1.40	1.89	1.85	-	Values taken from Pauling (1960) and used in Shrake & Rupley (1973). >C= can be either 1.50 or 1.85
Richards	1974	2.00	2.00	2.00	*	1.70	2.00	-	1.70	1.40	1.60	1.50	-	1.80	Minor modification of the (Lee & Richards, 1971), rationale not given. See original for discussion of aromtic carbon value.
Chothia	1975	1.87	1.87	1.87	1.76	1.76	1.50	1.65	1.65	1.40	1.40	1.40	1.85	1.85	From packing in amino acid crystal structures (Chothia, 1975)
Richmond & Richards	1978	1.90	1.90	1.90	1.70	1.70	1.70	1.70	1.70	1.40	1.40	1.40	1.80	1.80	No rationale given for values used (Richmond & Richards, 1978)
Gelin & Karplus	1979	1.95	1.90	1.85	1.90	1.80	1.75	1.70	1.65	1.60	1.70	1.60	1.90	1.90	Origin of values no specified (Gellin & Karplus, 1979)
Dunfield *et al.*	1979	2.13	2.23	2.38	2.10	1.85			1.75	1.56		1.62		2.08	Deconvolution of molecular crystal energies. Values are half of the heavy-atom separation at the min. of LJ 6-12 potential functions for symmetrical interactions. Used in Nemethy *et al.* (1983) and Dunfield *et al.* (1979)
ENCAD (derived)	1995	1.82	1.82	1.82	1.74	1.74	1.68	1.68	1.68	1.34	1.54	1.34	1.82	1.82	Set of radii derived in Gerstein *et al.* (1995), based solely on ENCAD MD potential (Levitt *et al.*, 1995). Based on LJ 6-12 where energy was 0.25 kBt (0.15 kcal/mol)
CHARMM (derived)	1995	1.88	1.88	1.88	1.80	1.80	1.40	1.40	1.40	1.38	1.53	1.41	1.56	1.56	Determined in same way as ENCAD set, but for CHARMM potential (Brooks *et al.*, 1983)
ProtOr	1999	1.88	1.88	1.88	1.76	1.61	1.64	1.64	1.64	1.42	1.46	1.42	1.77	1.77	Derived in Tsai *et al.* (1999) from analysis of the most common distances of approach of atoms in the Cambridge Structural Database.

Table 2. Radil Sets. All values in Å

cysteine). In this paper we perform a detailed analysis to see how sensitive these standard ProtOr volumes are to the various parameters that go into the calculations—e.g. radii set, structure set, etc. This analysis is in the spirit of formal mathematical sensitivity analysis (Rabitz, 1989), though less rigorous, reflecting the practical and empirical nature of protein packing calculations. As part of our analysis we make available on the web (at bioinfo.mbb.yale.edu/geometry and http://molmovdb.org) an extensive database of the many possible parameters and their associated volumes.

We calculate volumes of protein atoms by constructing polyhedra around them. Figure 2 outlines the parameters affecting this calculation. Briefly, these fall into five main groups: (i) the input structure set, (ii) the atom typing, (iii) the atom radii, (iv) the plane-positioning method and (v) the atom selection criteria. Each one of these parameters is changed while the others remain fixed. The fixed values for the parameters are detailed in Table 1 and are similar to the ones used previously (Gerstein *et al.*, 1995). For example, we use the ratio method (a variation of Richard's method B as described below) for all our calculations except in the cases where were are testing plane positioning methods. One instance of plane positioning is of special note, since in using it, the atom radii become irrelevant. Specifically, in the bisection plane-positioning method, an atom's radii has no influence on the placement of the plane, since the plane is placed midway between two atoms. The resulting polyhedra are true Voronoi constructs. Apart from this instance, the effect of the plane-positioning method can be separated from that of atom radii. However, atom typing and atom radii have strongly associated effects on the resulting protein volumes. This is the reasoning behind

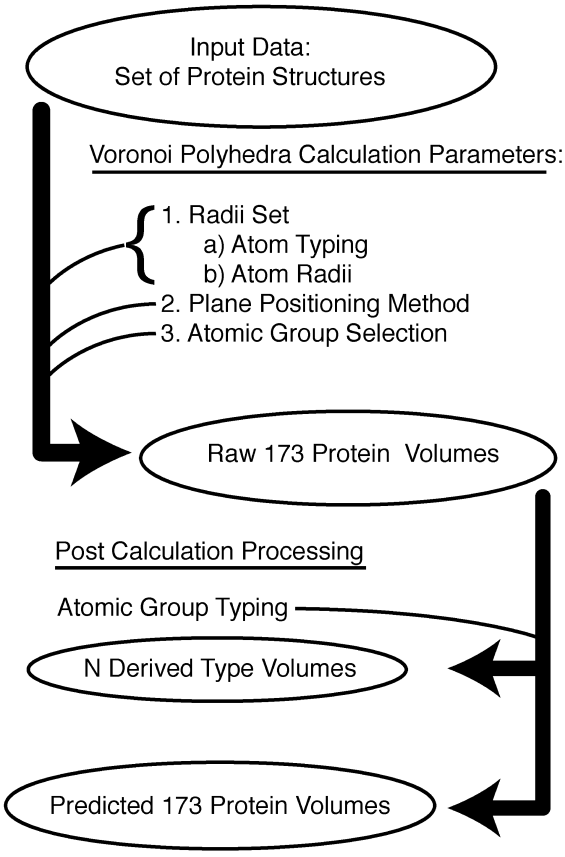

Fig. 2. Procedure and parameters for the Voronoi-like calculation. Note: we have 173 protein atoms from 21 amino acids. This latter number is 21 because we consider a reduced cysteine as another residue distinct from disulfide bonded cysteine. This adds six more protein atoms and one more amino acid.

how these parameters are displayed in Figure 2. In fact, most analyses usually combined these two parameters, discussing them together as a 'radii set'. Since the issue of atom typing is dealt with in a separate paper (Tsai *et al.*, 2001), we will only compare different radii sets in this work. Following this discussion, we will look at atom selection, a particularly problematic area with respect to using the polyhedra treatment. Finally, we also look into the effect that the type of structure set has on the resulting volumes. Previously, detailed work (Fleming and Richards, 2000) has shown that the characteristics of the structures in a set (i.e. protein size, secondary structure composition and amino acid content) effect packing calculations. This work does not go into such detail, but only attempts to look at various overall qualities of the structure set.

For our atom type notation (as in Table 1), the uppercase letter in the first register names the heavy atom (C, N, O and S for carbon, nitrogen, oxygen and sulfur). The number in the second register shows the number of covalent bonds the atom can make. The third register is always an H for hydrogen. The fourth register shows the number of hydrogen atoms connected to the heavy atom. Therefore, in our notation an sp^3 carbon with two hydrogens is C4H2 while a hydroxyl group is O2H1. Also, as explained below, an additional lowercase is used in the fifth register to describe the atom type: s, b or u (small, big or unique, respectively).

ANALYSIS OF THE EFFECT OF THE RADII SET

One might have thought that basic parameters like the van der Waals (VDW) radii of the various atoms would have fairly established and agreed upon values. Surprisingly, this is not the case. In Table 2, we document just some of the great number of different radii sets that have been proposed for proteins. Different radii sets will obviously have an effect on volume calculations. We try to give some indication of the overall sensitivity of the volume calculation to different sets of radii in Table 3, which shows residue volumes calculated from three of these radii sets: ProtOr (Tsai *et al.*, 1999), Chothia (Chothia, 1975) and Richards (Richards, 1974). In these calculations, all other parameters (structure set, atom selection method, etc.) remained fixed (see notes to Table 3). On average, the residue volumes from the ProtOr radii possess both lower deviations and lower volumes than results from the Chothia and Richards sets, although no residue volume is more than 5% different from another. The lower deviations suggest that the ProtOr set better fits the experimental data. The lower ProtOr volumes are a result of certain smaller atom volumes that dominate a residue's overall volume. For an atom, both of these are due not only to an atom's absolute radii, but also to its radii in relationship to the other radii in the set. As will be pointed out below, an accurate set of radii tends to minimize deviation found in side-chain volumes without causing an increase in main-chain atom volumes or their deviations.

To explain these effects in more detail, we produced histograms for four representative atom volumes: two main-chain atoms from alanine and two side-chain atoms, and present them in Figure 3. Looking at Table 3, the Richards radii set has the broadest range of radii (from 1.40 to 2.00 Å) of all three sets discussed here. From parts C and D of Figure 3, this range of radii produces side-chain atom distributions with the highest deviations and mean volumes. For main-chain atoms, alanine's carbon atom shows a similar distribution between the Richards and ProtOr radii sets, even though the radii for this atom is quite different (2.00 Å vs. 1.88 Å, respectively, Table 2). The oxygen atom of alanine also displays distributions of similar width between the Richards and ProtOr sets, but the Richards set's distribution possesses a lower mean.

Amino Acid	ProtOr Volume[c]	ProtOr SD	Chothia[a] Volume[c]	Chothia[a] SD	Richards[b] Volume[c]	Richards[b] SD
Gly	63.8	2.7	64.4	3.1	64.1	3.4
Ala	89.3	3.5	90.1	3.7	90.6	4.0
Val	138.2	4.8	139.0	4.5	140.4	4.7
Leu	163.1	5.8	164.0	5.3	165.7	5.5
Ile	163.0	5.3	163.6	5.5	165.7	5.7
Pro	121.3	3.7	122.7	4.6	123.8	4.8
Met	165.8	5.4	166.8	5.6	166.5	5.7
Phe	190.8	4.8	192.5	5.2	192.4	5.4
Tyr	194.6	4.9	195.9	5.1	197.9	5.5
Trp	226.4	5.3	229.8	5.5	228.5	5.6
Ser	93.5	3.9	94.4	3.8	97.2	4.4
Thr	119.6	4.2	119.8	4.7	122.6	5.1
Asn	122.4	4.6	125.3	5.0	126.0	5.5
Gln	146.9	4.3	148.5	5.3	149.5	5.6
Cys	112.8	5.5	113.9	4.2	115.3	4.5
Css	102.5	3.5	103.7	4.0	104.7	4.2
His	157.5	4.2	158.4	5.1	157.9	5.4
Glu	138.8	4.3	140.9	4.8	142.0	5.4
Asp	114.4	3.9	116.9	4.0	118.1	4.7
Arg	190.3	4.7	193.5	5.8	198.0	6.4
Lys	165.1	6.9	168.0	5.5	178.3	6.8

Table 3. Comparing output residue volumes from different radii/type sets. [a](Chothia, 1975), [b]((Richards, 1974)), [c]Volumes in Å^3. Calculations were done with the Standard PDB set and BL+ atom selection (see Table 5).

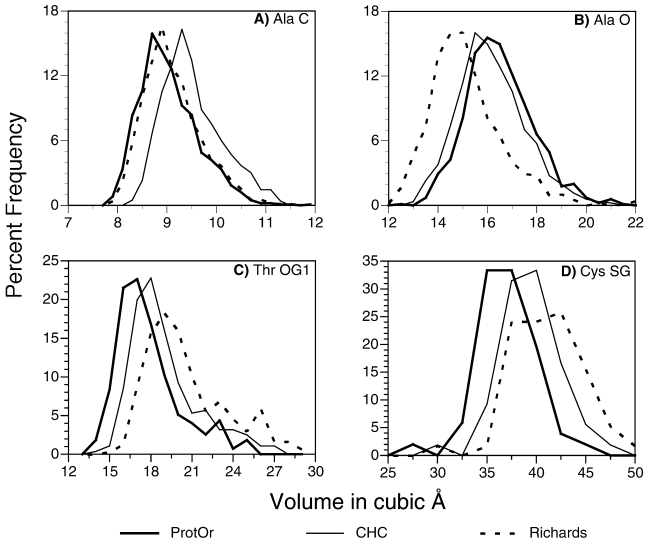

Fig. 3. Residue atom histograms. Normalized distributions of volumes for representative protein atoms are shown from Voronoi-like calculations using one of three different radii/typing sets (ProtOr, Chothia or Richards). All calculations used BL- selection, the Standard structure set, and the ratio method for plane-positioning. (a) Alanine's main-chain aliphatic carbon. (b) Alanine's main-chain oxygen. (c) Threonine's side-chain hydroxyl oxygen. (d) Cysteine's reduced sulfur.

The overall results of the Richards set in comparison to the ProtOr set are larger volumes and deviations primarily due to contributions of side-chain atoms.

Comparing the Chothia and ProtOr set is more difficult, since the radii are quite similar in the two sets (Table 3). Even such small differences produce different volumes. Although less obvious than the distributions from the Richards set, side-chain atom distributions produced by the Chothia radii are slightly shifted in comparison to those of the ProtOr set (Figure 3). In general, however, the ProtOr radii set produces slightly smaller residue volumes as shown in Table 3. This is due to the relationship between the radii in a set and how they partition space. In the Chothia set, the largest radius is given to the aliphatics at 1.87 Å (see Table 2). This radius is 6% larger than the next largest radius of 1.74 given to aromatic carbons. A similar comparison of the ProtOr set for the two largest atoms (also the aliphatics and aromatics) yields only a 4% increase. Since we are using the ratio method, the aliphatics in the Chothia set are partitioned more volume than in the ProtOr set. In summing the average values of atom types for residue volumes, the result causes slightly larger residue volumes for the Chothia set.

ANALYSIS OF THE EFFECT OF THE PLANE-POSITIONING METHOD

The simplest method for positioning the dividing plane between atoms is bisection. As noted above, bisection makes both the atom typing and radii set unnecessary since the plane is placed midway between two atoms. However, volumes of larger atoms are on average underestimated, and smaller ones overestimated. Two principal methods of re-positioning the dividing plane have been proposed to make the partition more physically reasonable: method B (Richards, 1974) and the radical-plane method (Gellatly and Finney, 1982). Both methods depend on the radii of the atoms in contact (R_1 and R_2) and the distance between the atoms D. They position the plane at a distance D_1 from the first atom. This distance is always set so that the plane is closer to the smaller atom. Method B is the simpler of the two and will be discussed in more detail here, since this method uses a linear proportionality

between two atoms' radii rather than the square used by the radical-plane method. For atoms that are covalently bonded, Method B divides the distance between the atoms according to their covalent-bond radii:

$$D_1 = D\frac{R_1}{R_1 + R_2}. \quad (1)$$

For atoms that are not covalently bonded, method B splits the remaining distance between them after subtracting their VDW radii:

$$D_1 = R_1 + \frac{D - (R_1+R_2)}{2}. \quad (2)$$

For separations that are not much different from the sum of the radii, the two formulas for method B (1) and (2) give essentially the same result. Consequently, it is worthwhile to try a slight simplification of method B, which we dub the 'ratio method'. Instead of using formula (1) for bonded atoms and formula (2) for nonbonded ones, one can use formula (2) in both. Doing this gives more consistent reference volumes (manifested in terms of smaller standard deviations about the mean). This is described more fully below.

We decided to measure the effectiveness of each plane-positioning method using the standard deviation of residue volumes. Chemically unreasonable partitioning of space will produce much more variable volumes for a given atom type, and this will manifest itself in larger standard deviations. The standard deviations about the mean for each of the 21 residue volumes are shown in Table 4. These were calculated using four different plane-positioning methods: bisection, method B, ratio and radical plane. The results came out as expected. All the chemically reasonable methods perform better than the bisection method. While one method does not show any significant advantage over the others (Table 4) we chose the ratio method over both method B and the radical plane method. The ratio method has only the most minor effect on the mean values (<0.5%). Since it is a simple proportion between two radii, the ratio method is a simple and chemically reasonable application of different radii. The ratio method has also been traditionally been used for these calculations (Richards, 1985). For these reasons, the ratio method was used for all the calculations reported here using a radii set. One drawback with repositioning the dividing plane with the ratio method is that the vertices are no longer exact and small volumes at each vertex are not assigned to any volume. For the ratio method, this vertex error has been calculated to be no more 0.2% of the total protein volume (Gerstein *et al.*, 1995). The radical plane method does not suffer from such vertex error, and this plane-positioning method would be more appropriate in cases where no vertex error was desired.

	Bisection	Original method B	Simple method B	Radical plane
Gly	5.8	5.0	5.0	5.1
Ala	5.4	4.5	4.4	4.5
Val	4.5	3.6	3.6	3.6
Leu	4.4	3.6	3.6	3.6
Ile	4.8	3.9	3.8	3.9
Pro	5.3	4.4	4.6	4.3
Met	4.9	4.1	4.2	4.1
Phe	3.9	3.3	3.2	3.3
Tyr	3.8	3.3	3.2	3.3
Trp	4.0	3.0	2.9	3.1
Ser	5.3	4.3	4.1	4.4
Thr	4.9	4.0	4.0	4.0
Asn	3.9	3.5	3.4	3.5
Gln	3.9	3.2	3.2	3.3
Cys	6.5	5.6	5.6	5.2
Css	5.5	4.5	4.5	4.6
His	4.8	3.7	3.7	3.8
Glu	4.3	3.9	3.8	3.9
Asp	4.4	3.8	3.7	3.9
Arg	3.9	3.2	3.3	3.2
Lys	4.1	2.8	2.8	3.0
Average	4.7	3.9	3.8	3.9

Table 4. The standard deviations about the mean volume for each of the 21 residue types, calculated with four different plane-positioning schemes. A bad plane-positioning scheme will, on average, misallocate volume, and this misallocation will, in turn, be manifest in the large variatons in polyhedra volume for chemically identical atoms—i.e. large standard deviations. Note that the simplified version of method B (ratio method) proposed here gives the lowest overall deviation

ANALYSIS OF THE EFFECT OF ATOM SELECTION CRITERIA

As discussed above, atom selection is necessary since constructing polyhedra around surface atoms remains an unsolved problem. The difficulty is that surface atoms have no neighbors towards what should be solvent, and polyhedra require neighbors for proper construction. To get around this problem, we try to choose atoms for which good polyhedra can be made. Table 5 shows total counts, volumes and standard deviations for each of the 18 ProtOr atom types using various criteria for selecting atoms to be included in the statistics. The analysis here is different from our previous work (Tsai *et al.*, 1999) in that it focuses on statistical rather than structural issues. However, the selection criteria are somewhat though not exactly similar. In the table, our selection criteria are ordered starting from least exclusive on the left to most exclusive on the right. The various criteria first exclude atoms at the surface

	Atom selection																				
	base			**B+**			**B-**			**BL-**			**BT**			**BL+**			**BD**		
	All atomic groups for which polyhedra can be constructed			Atomic groups with zero accessible surface area, waters included			Atomic groups with zero accessible surface area, waters excluded			Same as B- except those atomic groups touching non-protein atomic groups			Same as B+ less those touching the exposed atomic groups			Same as B+ except those atomic groups touching non-protein atomic groups			Same as BL+ less those atomic groups touching non-selected atomic groups		
ProtOr type	**Count**	**Vol.**	**SD**	**Count**	**Vol.**	**SD**	**Count**	**Vol.**	**SD**	**Count**	**Vol.**	**SD**	**Count**	**Vol.**	**SD**	**Count**	**Vol.**	**SD**	**Count**	**Vol.**	**SD**
C3H0b	12148	11.11	2.55	7802	9.78	0.72	6539	9.79	0.73	6090	9.79	0.74	4618	9.68	0.69	4255	9.67	0.68	1434	9.64	0.63
C3H0s	24588	9.05	1.08	20429	8.79	0.62	18341	8.79	0.64	17911	8.78	0.64	14011	8.69	0.59	13769	8.68	0.59	3647	8.73	0.58
C3H1b	5129	23.87	6.65	3668	21.40	1.88	3074	21.39	1.88	2773	21.40	1.87	2376	21.38	1.89	2363	21.38	1.89	1046	21.33	1.84
C3H1s	5902	23.95	7.79	3874	20.57	1.79	2941	20.56	1.82	2627	20.56	1.81	1973	20.43	1.78	1938	20.41	1.77	767	20.52	1.71
C4H1b	10248	16.05	4.12	7440	14.47	1.23	5974	14.47	1.23	5719	14.46	1.23	4634	14.42	1.22	4579	14.41	1.22	1947	14.37	1.17
C4H1s	20389	14.66	3.25	14128	13.31	0.99	10882	13.27	1.01	10455	13.26	1.00	8358	13.17	0.97	8280	13.17	0.96	3274	13.17	0.90
C4H2b	4391	30.62	11.82	2411	24.43	2.20	1712	24.41	2.19	1564	24.46	2.20	1162	24.28	2.14	1152	24.25	2.13	559	24.13	2.06
C4H2s	21531	29.18	10.06	10234	23.48	1.97	6421	23.52	2.01	5927	23.54	2.00	4010	23.30	1.94	3948	23.29	1.94	1989	23.33	1.88
C4H3u	12891	41.84	11.13	7675	36.87	3.05	5711	36.96	3.06	5241	36.98	3.02	4179	36.94	2.99	4160	36.94	2.99	1890	36.74	2.99
N3H0u	1119	9.14	1.92	948	8.70	0.73	848	8.80	0.74	830	8.79	0.72	501	8.58	0.66	495	8.57	0.65	107	8.48	0.56
N3H1b	2047	20.18	8.98	1212	15.74	2.02	769	15.76	2.42	648	16.07	1.98	449	15.75	1.73	437	15.73	1.70	228	15.31	2.09
N3H1s	25551	14.87	3.91	20712	13.72	1.06	16057	13.77	1.19	15354	13.76	1.18	11224	13.53	1.00	11180	13.53	1.00	3282	13.74	0.99
N3H2u	2276	33.13	15.23	878	22.82	2.33	406	23.05	2.84	345	23.03	2.89	179	22.08	2.13	178	22.07	2.13	142	22.54	2.10
N4H3u	401	41.77	22.36	88	21.32	3.77	26	20.20	6.51	19	23.07	3.40	6	21.03	1.29	6	21.03	1.29	9	21.56	2.61
O1H0u	28430	20.84	10.21	18670	16.07	1.45	12545	16.14	1.55	11781	16.15	1.51	8865	15.92	1.28	8847	15.92	1.28	2759	16.07	1.30
O2H1u	3289	25.29	12.66	1842	18.08	1.93	910	18.63	2.38	780	18.67	2.34	479	18.10	1.87	477	18.09	1.86	369	17.82	1.49
S2H0u	843	32.05	9.04	610	28.64	2.84	507	28.68	2.87	461	28.80	2.61	354	28.80	2.67	352	28.79	2.67	115	28.16	3.28
S2H1u	167	35.75	8.92	119	33.39	5.38	94	33.32	4.85	55	36.23	2.62	47	35.93	2.44	47	35.93	2.44	34	33.92	3.98

Table 5. Comparision of atom selection methods. Volumes in $Å^3$. The fluctuation in polyhedron volume over the sample ('S.D.' column in Tables 5 and 7) is expressed in terms of the standard deviation of all the individual measurements, taken together, as a percentage of the mean. To get the error in the mean (i.e. the standard deviation of a distribution of mean values), it is necessary to divide by the square root of the number of measurements N. Performing this operation for the mainchain carbonyl carbon yields an expected error in the mean of only 0.08%.

and then progressively more and more atoms towards the core, since the core has been shown to give more regular volumes (Chothia, 1974; Finney, 1975; Richards, 1974).

base The base method of atom selection considers all protein atoms except those for which volumes cannot be calculated (usually because of open-ended polyhedra).

B The next level of filtering removes atoms that are exposed to solvent. Determination of whether an atom is exposed to water was done using the conventional Lee and Richards accessible surface area (1971).

BL At this level of selection, those atoms that touch ligands or any non-protein atom are not considered in addition to surface atoms.

BT This level of selection starts with removing all exposed atoms as in the B selection method, but also does not consider those atoms that are neighbors or directly touching an exposed atom.

BD The most stringent selection used in this work removes all exposed atoms and all atoms touching non-protein atoms. In addition, any atom neighboring an atom touching a non-protein atom is not used.

For some of the selection schemes described above, the inclusion of the crystallographic waters influenced the calculation of volumes. For these, a '+' is added when crystallographic waters are used, and a '−' is added when they are not. As shown in the Table 5, increasing the degree of exclusion decreases the volumes of each of the types as well as the counts and standard deviation. This raises an important point. To obtain accurate protein volumes, the selection method needs to strike a balance between minimizing the standard deviation and using a large enough population of atoms to average over. This is not an issue for most atoms such as the aliphatic and aromatic carbons. Populations for these types of atoms never decrease below 500 counts, even in our strictest atom selection regime. However, the small sample size of lysine's charged nitrogen and cysteine's reduced sulfur (the N4H3u and S2H1u atom types, respectively) were of some concern. We, therefore, discuss the selection sets in numbers of counts in these marginal atoms. As the selection becomes more stringent, we see an expected decrease in counts and standard deviations for both N4H3u and S2H1u atom types. One notable exception is the strict, BD selection. It decreases most counts significantly without a simultaneous reduction in deviations. Also, surprisingly, this level of selection actually considers a higher number of N4H3u groups than previous selection methods, but the deviation also increases. The selection method that balances numbers with a small deviation is either the BT or BL+ methods.

Set	Number	Identifier
Standard	130	135l, 1aaj, 1aap, 1ake, 1arb, 1bbh, 1bp2, 1ccr, 1cdp, 1cmb, 1cpc, 1crn, 1cse, 1ctf, 1cus, 1dfn, 1dr1, 1eco, 1ezm, 1fkf, 1fus, 1fxd, 1gct, 1gd1, 1gpr, 1hbg, 1hel, 1hne, 1ifc, 1igd, 1lmb, 1lz1, 1lz3, 1mba, 1mbd, 1ofv, 1omd, 1paz, 1pgx, 1pk4, 1plc, 1ppn, 1ppt, 1ptx, 1rcf, 1rdg, 1rms , 1rop, 1rpg, 1rpo, 1rro, 1sar, 1sgt, 1snc, 1st3, 1thm, 1ubq, 1ycc, 256b, 2act, 2alp, 2apr, 2aza, 2cba, 2ccy, 2cdv, 2cpp, 2ctc, 2cyp, 2er7, 2fb4, 2fcr, 2fx2, 2gbp, 2hhb, 2ihl, 2ltn, 2mcm, 2mhr, 2msb, 2ovo, 2por, 2prk, 2rhe, 2rn2, 2sga, 2sn3, 2trx, 2utg, 2wrp, 2zta, 3app, 3b5c, 3bcl, 3c2c, 3cla, 3dfr, 3ebx, 3est, 3fxn, 3grs, 3lzm, 3rp2, 3sgb, 451c, 4dfr, 4enl, 4icb, 4ins, 4ptp, 5cpa, 5cyt, 5p21, 5pal, 5pti, 5rub, 5rxn, 5tim, 6ebx, 6rlx, 6rxn, 6xia, 7aat, 7rsa, 8dfr, 8fab, 8rxn, 9pti, 9rnt, 9wga
SCOP	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
NMR	125	1aab, 1aaf, 1aca, 1acp, 1afp, 1ahd, 1ale, 1alf, 1bbo, 1bus, 1bw3, 1bw4, 1cdb, 1cdn, 1cis, 1clb, 1crp, 1crq, 1crr, 1csy, 1csz, 1ctl, 1dhm, 1erg, 1erh, 1fht, 1fkr, 1fks, 1fkt, 1ftz, 1gb1, 1gbr, 1gfc, 1gfd, 1hcc, 1hdn, 1hme, 1hmf, 1hom, 1hrq, 1hrr, 1hsm, 1hsn, 1hue, 1hum, 1hun, 1il8, 1iml, 1irp, 1kb7, 1kb8, 1ldl, 1ldr, 1lip, 1lpt, 1mbe, 1mbf, 1mbg, 1mbj, 1mbk, 1mef, 1ncp, 1neh, 1neq, 1ner, 1nhm, 1nhn, 1nil, 1nim, 1nmf, 1nmg, 1noe, 1odp, 1odq, 1odr, 1oef, 1oeg, 1pan, 1pao, 1pcp, 1pdc, 1pih, 1pij, 1pmc, 1pog, 1pra, 1prr, 1prs, 1pse, 1psf, 1qwe, 1qwf, 1rht, 1rip, 1rod, 1rpv, 1san, 1sap, 1srl, 1srm, 1stu, 1sxl, 1tam, 1tiv, 1tvs, 1tvt, 1ums, 1umt, 1utr, 1vnd, 1zer, 2abd, 2bus, 2gb1, 2gva, 2gvb, 2hid, 2hmx, 2hoa, 2igg, 2igh, 2il8, 2ptl, 2znf, 3ci2
Current	69	116l, 1act, 1alp, 1alr, 1anh, 251c, 156b, 1apd, 2bcl, 1abk, 1abp, 1abx, 1afg, 1ace, 1afn, 1ak3, 1asi, 1aza, 1baa, 1bjl, 2grs, 1cab, 1cae, 1cd4, 1ci2, 1cpk, 1cln, 1dhb, 1dri, 1eip, 1end, 7atc, 1fnr, 1gap, 1gbp, 1gcr, 1gmf, 1gn5, 2hvt, 1gsr, 1gyi, 1hft, 1hid, 1hmg, 1hmx, 1lrd, 3fab, 1mev, 1omf, 1ora, 1pab, 1pel, 1pgk, 1phy, 1ptc, 1r04, 1r1e, 1rsl, 1sod, 1srt, 1tbs, 1tct, 1trt, 1yhx, 2adk, 1vaa, 1ts1, 1ada
Obsolete	69	1abe, 1cdh, 1eri, 1fnb, 1lmb, 216l, 256b, 2abk, 2abx, 2ace, 2act, 2ada, 2afg, 2afn, 2ak3, 2alp, 2alr, 2anh, 2apd, 2asi, 2aza, 2baa, 2cab, 2cae, 2ci2, 2cpk, 2cyh, 2dhb, 2dri, 2eip, 2end, 2gmf, 2gn5, 2gsr, 2gyi, 2hft, 2hid, 2hmg, 2hmx, 2mev, 2omf, 2ora, 2pab, 2pel, 2phy, 2ptc, 2r04, 2rsl, 2sod, 2srt, 2tbs,2tct, 2trt, 2ts1, 2vaa, 2yhx, 351c, 3adk, 3bcl, 3bjl, 3cln, 3gap, 3gbp, 3grs, 3hvt, 3pgk, 4gcr, 5at1, 7fab

Table 6. Structure Sets

Since the BL+ method is slightly more accurate without a great loss in numbers of counts, we chose this method of atom selection for our standard set of volumes.

ANALYSIS OF THE EFFECT OF STRUCTURE SETS

Different sets of structures

The final sets of standard volumes that we calculate are obviously contingent on the set of structures input into the calculation. To ascertain the extent of this effect, we created a number of different sets of structures. PDB identifiers for each set are shown in Table 6.

Standard. The Standard set of proteins chosen for this study are the set of 119 used previously (Harpaz *et al.*, 1994), augmented to a total of 130 proteins by adhering to the same criteria used to choose the first 119 structures: resolution between 1.0 and 1.9 Å, *R*-factor less than 20%, and good stereochemistry.

High/low. To test the effect of the PDB set on resulting protein volumes, we split these structures in half, based on resolution. The cutoff was at 1.63 Å. This procedure produces high and low resolution structure sets, each containing 65 structures.

NMR. The NMR set consists of 125 structures. We ran the calculation over all of the structures to get average values for that protein.

Current/obsolete. For a more rigorous comparison, we constructed a structure set from the PDB and

ProtOr atom type	PDB sets [a]													
	SCOP		Standard		High		Low		NMR		Current		Obsolete	
	Vol. [b]	SD	Vol. [b]	SD	Vol. [b]	SD	Vol. [b]	SD	Vol. [b]	SD	Vol. [b]	SD	Vol. [b]	SD
C3H0b	9.64	0.72	9.67	0.68	9.65	0.68	9.68	0.69	9.53	1.05	9.78	0.79	9.83	0.86
C3H0s	8.66	0.58	8.68	0.59	8.65	0.57	8.70	0.60	8.65	0.80	8.77	0.69	8.84	0.76
C3H1b	21.33	1.87	21.38	1.89	21.36	1.85	21.39	1.91	19.40	2.73	21.26	2.11	20.96	2.30
C3H1s	20.45	1.76	20.41	1.77	20.27	1.72	20.50	1.80	18.48	2.78	20.42	2.02	20.43	2.21
C4H1b	14.35	1.35	14.41	1.22	14.38	1.20	14.43	1.23	13.89	1.55	14.40	1.48	14.42	1.59
C4H1s	13.14	0.97	13.17	0.96	13.20	0.94	13.15	0.97	13.20	1.27	13.11	1.11	13.18	1.20
C4H2b	24.14	2.07	24.25	2.13	24.11	1.95	24.33	2.21	20.48	5.89	24.26	2.43	24.07	2.76
C4H2s	23.17	2.35	23.29	1.94	23.28	1.96	23.29	1.93	19.13	6.40	23.14	2.23	22.92	2.46
C4H3u	36.84	3.24	36.94	2.99	36.93	3.00	36.94	2.98	30.38	8.26	36.43	3.75	35.76	3.95
N3H0u	8.62	0.59	8.57	0.65	8.60	0.70	8.56	0.63	9.01	0.92	8.70	0.72	8.79	0.79
N3H1b	15.65	1.55	15.73	1.70	15.55	1.48	15.80	1.79	15.19	2.44	15.99	2.00	16.25	2.28
N3H1s	13.54	0.99	13.53	1.00	13.52	0.97	13.53	1.01	13.67	1.69	13.64	1.30	13.72	1.51
N3H2u	22.61	2.36	22.07	2.13	22.12	2.22	22.04	2.09	15.11	5.15	22.48	2.72	22.35	3.06
N4H3u	21.56	1.28	21.03	1.29	20.30	0.55	21.76	1.40	17.80	5.06	23.85	3.00	23.06	2.42
O1H0u	15.91	1.29	15.92	1.28	15.87	1.23	15.94	1.30	15.17	2.19	15.95	1.74	15.98	1.98
O2H1u	18.11	1.78	18.09	1.86	18.10	1.97	18.09	1.79	17.48	2.81	18.67	2.55	19.09	3.16
S2H0u	29.29	2.68	28.79	2.67	28.66	2.68	28.90	2.66	26.25	3.44	29.78	3.26	29.94	3.49
S2H1u	36.82	3.48	35.93	2.44	37.15	2.46	35.71	2.38	30.71	5.89	35.80	3.35	35.86	3.46

Table 7. Comparision of structure sets. [a]Explanation of the various PDB sets discussed with more detail in the methods and Table 5B. Calculations used the ProtOr radii/type set and BL+ atom selection (see Table 5). [b]Volumes in $Å^3$.

then found their obsolete counterparts from the Archive of Obsolete PDB structures (a joint initiative of the San Diego Supercomputer Center and the Protein Data Bank that can be found at: http://www.sdsc.edu/PDBobsolete). If more than one obsolete structure existed, we always chose the earliest one solved. Overall, we have 69 current and obsolete structures.

SCOP. Finally, we derive a new set of 87 structures to calculate the final protein volumes that are better representative of different protein environments. Based on a 1.75 Å resolution cutoff, these structures were chosen from a larger list of structures that contained the best representative of a SCOP, the Structural Classification of Proteins database, classified domain (Murzin *et al.*, 1995). Hence, our set of 87 structures is named the SCOP set.

Comparisons of the structure sets: higher quality, smaller volumes

We show derived atom type volumes using different PDB sets in Table 7. All parameters were the same except for the set of structures that the calculation ran over (see notes to Table 7). One of the main factors in building a structure set is the quality of the data. Split from the Standard set (see above), the high and low sets look at the effect of resolution. The average volumes in the low-resolution set are usually larger than those in the high-resolution set by 1% on average. While the average standard deviation increases by 11% on average in the low-resolution set, this is primarily due to a large change in the N4H3u deviation. Without its contribution, the increase drops to 5%. Comparison of the standard and NMR sets shows an even greater change. On average, we find a 4% decrease in volume and 31% increase in standard deviation, which is in direct contrast to the previous comparison of high- and low-resolution structure sets.

So far, these comparisons of PDB sets are problematic in that all of these sets contain different structures. A more rigorous comparison would be to use two sets containing the same structures but possessing different resolutions. This is the premise behind the current and obsolete sets. The current set's volumes differ by an average increase of 1% over the obsolete set, although of the 18 atom types, a majority of 13 show a decrease in volume from obsolete to current. The standard deviations a more uniform decrease in comparing obsolete to current. As expected, these results indicate that higher resolution structures generally produce smaller volumes and smaller standard deviations.

Resolution and standard deviation are not the only factors influencing the quality of the resulting volumes. Because the volume of an atom depends upon its neighbors, a set of structures should adequately represent the many different protein environments and avoid any redundancy, such as homologs. This inspired us to find a new structure set consisting of structures that are as distinct as possible from each other. Out of the search, we obtained the SCOP set, as described above. Comparing results from the SCOP set to the standard set, we find that the volumes change by

only 0.7% on average. Standard deviations change on average by 7.3%. In both volume and standard deviation, the greatest difference occurs in atoms of side-chains and of those, the greatest difference occurs in the ones towards the ends, i.e. the reduced sulfur atom S2H1u and the end methyl group C4H3u. Even with these differences, we are confident that these volumes using the SCOP set are a more accurate representation of protein atom volumes.

CONCLUSION

The volumes, radii, typing and associated parameters summarized in Table 1 make up the ProtOr volume and parameter set. In deriving this set, we have found that selecting atoms using the BL+ method properly balances a low standard deviation with an adequate sample population. Even though waters are included, the atoms touching them are removed with this selection method. The choice of structures to run over is also important. Higher-resolution sets generally decrease the volume and standard deviation, but for representative volumes, the structure set should consist of as many different atom environments as possible, such as the SCOP set used here.

One outcome of this work was the development of a set of standard residue volumes that would be generally useful in calculations of protein properties. To such an end, we present volumes for amino acid residues computed with the ProtOr set in Table 8. Two sets of volumes are shown, taken from different points in the overall calculation (see Figure 2). Basically, the two differ in the number of volumes used to compute the residue volumes. The output set uses the raw 173 protein atom volumes, while the predicted set uses *n* derived atom type volumes collapsed from the raw 173 atom volumes. Therefore, the former volumes are most likely more accurate, and for future reference, we consider this set of residue volumes the official ProtOr residue volumes. In conclusion, these volumes show just one application of the ProtOr set, and we hope to have shown the ProtOr set's general usefulness and accuracy in calculations requiring protein atom radii.

Amino acid	ProtOr 98 [a] Volume / Å^3	Predicted [b] Volume / Å^3
Gly	63.8	62.4
Ala	89.3	89.3
Val	138.2	139.2
Leu	163.1	162.4
Ile	163.0	163.5
Pro	121.3	119.3
Met	165.8	164.8
Phe	190.8	189.9
Tyr	194.6	193.7
Trp	226.4	225.0
Ser	93.5	92.6
Thr	119.6	120.5
Asn	122.4	122.9
Gln	146.9	146.1
Cys	112.8	112.4
Css	102.5	104.8
His	157.5	156.6
Glu	138.8	139.3
Asp	114.4	116.1
Arg	190.3	191.8
Lys	165.1	166.7

Table 8. Residue volumes. We consider reduced cysteine (CYS) as dinstinct from disulfide bonded cysteine (CSS). [a]These are raw output residue volumes obtained from summing the residues' respective atom volumes from the raw 173 atom volumes (see Figure 2) output after a polyhedra calculaion using the ProtOr atom type set, the Best pdb set, and BL+ atom selection (see Table 5). [b]As explained within Figure 2, predicted residue volumes obtained from the ProtOr derived atom type volumes.

SOURCE CODE AND PARAMETER DATABASE, AVAILABLE ON THE WEB

We make available a general code base for geometric calculations on macromolecular structures. This includes: (1) code and executables for calculating Voronoi-like polyhedra and Delaunay triangulations and (2) programs to calculate related geometric quantities. We also make available an extensive collection (i.e. database) of geometric parameters associated with the calculations. This includes standard volumes, typing schemes, radii sets, lists of structures, etc, comprising more than 500 distinct schemes in over 1000 files. These can be retrieved by emailing Mark.Gerstein@yale.edu or going to http://bioinfo.mbb.yale.edu/geometry and http://molmovdb.org.

ACKNOWLEDGEMENTS

J.T. would like to thank the National Institutes of Health (grant number GM41455) and the NSF Biological Informatics Fellowship) for support. M.G. would like to thank the National Science Foundation for support (grant number DBI-9723182). We also appreciate the kind support and suggestions from Cyrus Chothia, Michael Levitt, Fred Richards and David Baker.

REFERENCES

Abola,E.E., Sussman,J.L., Prilusky,J. and Manning,N.O. (1997) Protein data bank archives of three-dimensional macromolecular structures. *Meth. Enzymol.*, **277**, 556–571.

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

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

Bondi,A. (1964) van der Waals volumes and radii. *J. Phys. Chem.*, **68**, 441–451.

Chothia,C. (1974) Hydrophobic bonding and accessible surface area in proteins. *Nature*, **248**, 338–339.

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

David,C.W. (1988) Voronoi polyhedra as structure probes in large molecular systems. *Biopolymers*, **27**, 339–344.

Dunbrack,Jr.,R.L. (1999) Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. *Proteins*, **Suppl 3**, 81–87.

Finney,J.L. (1975) Volume occupation, environment and accessibility in proteins. the problem of the protein surface. *J. Mol. Biol.*, **96**, 721–732.

Finney,J.L. (1978) Volume occupation, environment, and accessibility in proteins. environment and molecular area of RNase-S. *J. Mol. Biol.*, **119**, 415–441.

Fleming,P.J. and Richards,F.M. (2000) Protein packing: dependence on protein size, secondary structure and amino acid composition. *J. Mol. Biol.*, **299**, 487–98.

Gellatly,B.J. and Finney,J.L. (1982) Calculation of protein volumes: an alternative to the Voronoi procedure. *J. Mol. Biol.*, **161**, 305–322.

Gerstein,M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.*, **3**, 497–512.

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

Gerstein,M. and Krebs,W. (1998) A database of macromolecular motions. *Nucleic Acids Res.*, **26**, 4280–4290.

Gerstein,M., Sonnhammer,E.L. and Chothia,C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078.

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.

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

Hubbard,S.J. and Argos,P. (1995) Detection of internal cavities in globular proteins. *Protein Engg*, **8**, 1011–1015.

Janin,J. (1979) Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.

Janin,J. and Chothia,C. (1990) The structure of protein-protein recognition sites. *J. Biol. Chem.*, **265**, 16027–16030.

Koehl,P. and Delarue,M. (1997) The native sequence determines side-chain packing in a protein, but does optimal side-chain packing determine the native sequence? *Pac. Symp. Biocomput.*, 198–209.

Krebs,W.G. and Gerstein,M. (2000) The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res.*, **28**, 1665–1675.

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

Lee,C. and Levitt,M. (1997) Packing as a structural basis of protein stability: understanding mutant properties from wildtype structure. *Pac. Symp. Biocomput.*, 245–255.

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.

Liang,J., Edelsbrunner,H., Fu,P., Sudhakar,P.V. and Subramaniam,S. (1998a) Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins*, **33**, 1–17.

Liang,J., Edelsbrunner,H., Fu,P., Sudhakar,P.V. and Subramaniam,S. (1998b) Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins*, **33**, 18–29.

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

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

Rabitz,H. (1989) Systems analysis at a molecular scale. *Science*, **246**, 221–226.

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. *Meth. Enzymol.*, **115**, 440–464.

Thorton,J.M. (1992) Lessons from analyzing protein structures. *Curr. Opin. Struct. Biol.*, **2**, 888–894.

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.

Tsai,J., Voss,N. and Gerstein,M. (2001) Determining the minimum number of types necessary to represent the sizes of protein atoms. *Bioinformatics*, **17**, 949–956.

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