Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Accurate calculation of the density of proteins

Michael L. Quillin and Brian W. Matthews*

Institute of Molecular Biology, Howard Hughes Medical Institute and Department of Physics, 1229 University of Oregon, Eugene, Oregon 97403-1229, USA

Correspondence e-mail: brian@uoxray.uoregon.edu

Received 7 February 2000
Accepted 3 May 2000

On the basis of theoretical calculations, Andersson & Hovmöller have recently suggested that the long-established value of 1.35 g cm^{-3} for the mean density of proteins should be revised to 1.22 g cm^{-3} [Andersson & Hovmöller (2000), *Acta Cryst.* D**56**, 789–790]. To substantiate their assertion, these authors used the Voronoi algorithm to calculate the mean atomic volume for 30 representative protein structures. The Voronoi procedure requires that atoms of interest be bounded on all sides by other atoms. Volume calculations for surface atoms that are not surrounded or are only sparsely surrounded by other atoms either are not possible or may be unreliable. In an attempt to circumvent this problem, Andersson & Hovmöller rejected atoms with calculated volumes that were indeterminate or were greater than 50 Å^3. In the present study, it is shown that this criterion is not sufficiently restrictive to ensure accurate volume determinations. When only strictly buried atoms are included in the volume calculations using the Voronoi algorithm, the mean density is found to be 1.47 $\pm$ 0.05 g cm^{-3}. In addition, an alternate procedure based on the Connolly algorithm that permits all protein atoms to be included in volume calculations gives 1.43 $\pm$ 0.03 g cm^{-3} for the mean density of the same set of proteins. The latter two calculated values are mutually consistent and are in better agreement with the experimental value.

1. Introduction

In addition to being an important biophysical property in its own right, the mean density of proteins is a useful quantity for X-ray structure analysis. Estimates of the solvent content within protein crystals are required for many crystallographic techniques, including the determination of the number of molecules per unit cell (Matthews, 1968) and phase improvement by solvent flattening (Wang, 1985; Leslie, 1987). These estimates rely to a large extent upon an accurate value for the density of proteins. The generally accepted value of approximately 1.35 g cm^{-3} is derived from the partial specific volumes of proteins measured using hydrodynamic (Kuntz & Kauzmann, 1974; Squire & Himmel, 1979) and adiabatic compressibility (Gekko & Noguchi, 1979) methods. Recently, the use of this value has been brought into question on the basis of molecular-volume calculations conducted by Andersson & Hovmöller (1998, 2000). By applying the Voronoi algorithm (Voronoi, 1905; Richards, 1974; Finney, 1975; Gerstein *et al.*, 1995) to a set of 30 protein crystal structures, they obtained a value of 1.22 $\pm$ 0.02 g cm^{-3} for the mean density of proteins.

Andersson & Hovmöller are not the first to report a discrepancy between experimental and calculated values for the density of proteins. Preliminary calculations by Kauzmann *et al.* (1974) gave values of 1.09–1.25 g cm^{-3} for the densities of nine proteins for which structures had been determined at that time. The corresponding experimental values from solution studies were 1.33–1.42 g cm^{-3}. With the development of procedures for using the Voronoi algorithm (Voronoi, 1905) to calculate molecular volumes of proteins analytically (Richards, 1974; Finney, 1975), however, the agreement between experimental and theoretical values for protein densities improved dramatically. One disadvantage of using the Voronoi algorithm to determine the volumes of protein atoms is that owing to the absence of surrounding atoms, the Voronoi polyhedra that are generated for atoms at or near the surface of the protein often are not closed or have overly large volumes. In their calculations, Richards and Finney addressed this limitation either by surrounding the protein with a cubic lattice of water molecules (Richards, 1974) or by placing water molecules at tetrahedral loci on the surface of the protein (Finney, 1975). More recently, Gerstein and coworkers

Table 1
Molecular weights, volumes and densities for representative proteins calculated using the Voronoi and Connolly methods.

		Values from Voronoi method						Values from Connolly method		
		Buried atoms			Exposed atoms†			All atoms		
Protein name	PDB entry‡	Molecular weight (g mol^{-1})	Volume (Å^3)	Density (g cm^{-3})	Molecular weight (g mol^{-1})	Volume (Å^3)	Density (g cm^{-3})	Molecular weight (g mol^{-1})	Volume (Å^3)	Density (g cm^{-3})
Alcohol dehydrogenase	2ohx	37948	44986	1.40	28358	55236	0.85	77550	93059	1.38
Aldolase (2.0 Å)	2fua§	9259	10538	1.46	9195	17541	0.87	23127	27094	1.42
Aldolase (2.67 Å)	3fua§	9647	10959	1.46	8505	16424	0.86	22676	26300	1.43
Carbonic anhydrase	4cac	13522	15509	1.45	9806	19316	0.84	28757	33866	1.41
Carboxypeptidase A	2ctb§	16598	19433	1.42	12174	22518	0.90	34489	40613	1.41
Chymotrypsinogen	1chg	9979	11073	1.50	9154	17908	0.85	23457	27227	1.43
Cytochrome b_{562}	256b§	3775	4152	1.51	5234	9883	0.88	11780	13304	1.47
Cytochrome *c*	2ycc	4084	4311	1.57	5149	9806	0.87	12062	13450	1.49
Cytochrome P450	8cpp	21083	24726	1.42	16927	32958	0.85	45590	53808	1.41
Enolase	1one	47211	55319	1.42	34312	66125	0.86	93353	110803	1.40
F1 ATPase	1bmf	149271	172049	1.44	137735	284372	0.80	322329	383167	1.40
γ-B crystallin	4gcr§	8776	9795	1.49	7272	13714	0.88	20972	23743	1.47
Glucose oxidase	1gal	31422	35978	1.45	22941	43030	0.89	63063	73789	1.42
Glycolate oxidase	1gox	16442	19888	1.37	15531	29542	0.87	38379	45660	1.40
Hemoglobin (1.5 Å)	1thb	26443	28771	1.53	26383	54388	0.81	62023	71527	1.44
Hemoglobin (2.0 Å)	1gbv	24768	27479	1.50	27545	56612	0.81	61930	72310	1.42
Hemoglobin V	2lhb§	5306	5899	1.49	7174	13743	0.87	16280	18813	1.44
Insulin	1bph§	2016	2234	1.50	1563	2821	0.92	5679	6371	1.48
L-Arabinose binding protein	6abp	15052	17534	1.43	11666	21856	0.89	32917	38805	1.41
Myohemerythrin	2mhr§	5075	5557	1.52	5303	9414	0.94	13782	15746	1.45
Neuraminidase	2sim	20429	23824	1.42	14985	27989	0.89	41953	49438	1.41
Pyruvate kinase	1pkm	23534	26519	1.47	23990	47810	0.83	56803	66072	1.43
Repressor of primer	1rop	1623	1674	1.61	2766	4607	1.00	6405	7002	1.52
Retinol binding protein	1hbp	6874	7663	1.49	8700	16365	0.88	20055	22629	1.47
Rubisco	1rsc	28997	33373	1.44	23814	46383	0.85	62834	73480	1.42
Satellite panicum mosaic virus	1stm	31362	36610	1.42	28267	53095	0.88	75483	88668	1.41
Superoxide dismutase	3sdp	16875	18515	1.51	16467	32202	0.85	40983	47207	1.44
Triosephosphate isomerase	4tim	24593	29195	1.40	20729	39379	0.87	53427	63621	1.39
Trypsin	2trm	10262	11632	1.46	9024	16397	0.91	23796	27386	1.44
Xylose isomerase	2gyi	34028	38831	1.46	36893	71176	0.86	85391	98815	1.43
Mean				1.47			0.87			1.43
Standard deviation				0.05			0.04			0.03

† For the exposed atoms, the molecular weight, volume and density are for all atoms with non-zero solvent exposure that permit calculation of a Voronoi volume (see text). ‡ Original references are given by Andersson & Hovmöller (1998). § These entries contain alternate conformations or two molecules in the asymmetric unit which are represented by average values in this table.

extended this strategy by taking advantage of crystallographically observed water molecules (Gerstein & Chothia, 1996) and by using molecular-dynamics simulations to position water molecules near the surface of the protein (Gerstein *et al.*, 1995). In their characterization of standard atomic volumes obtained from crystal structures, Pontius *et al.* (1996) adopted the simple yet effective strategy of rejecting Voronoi volumes of all atoms which have any exposure to solvent. In their calculation of protein densities, Andersson & Hovmöller (1998) were aware of the caveats of the Voronoi procedure and attempted to circumvent the problem by excluding atoms for which the volume was either undefined (*i.e.* the Voronoi polyhedron was not closed) or had a value greater than 50 Å^3. We show that these criteria are not sufficiently restrictive. Using the approach of Pontius *et al.* (1996), we obtain a substantially higher value for the mean density of the same 30 protein structures analyzed by Andersson & Hovmöller (1998). In addition, by using the rolling-probe algorithm of Connolly (1993) to calculate the volumes of all atoms within these structures, we obtain values for protein density which agree closely with the revised values calculated with the Voronoi algorithm and which can be reconciled with experimentally determined protein densities.

2. Methods

To provide a strict comparison between protein densities calculated using different procedures, the 30 crystal structures of 28 different proteins selected by Andersson & Hovmöller (1998) were also analyzed in the present investigation. Coordinates for each structure were obtained from the Protein Data Bank (Bernstein *et al.*, 1977). To eliminate artifacts arising from non-protein atoms, all cofactors, ions, inhibitors and water molecules were removed. Explicit H atoms and atoms with zero occupancy, if present, were also discarded. Proteins which exhibited alternate conformations (1bph, 2ctb, 2fua, 2mhr, 3fua and 4gcr) or sequences (2lhb) were segregated, treated independently and the results averaged. Two molecules related by non-crystallographic symmetry in one structure (256b) were likewise evaluated separately and averaged.

Molecular volumes were determined using two independent methods. The first method employed the Connolly algorithm as implemented in the *MSP* package (Connolly, 1993), using the default van der Waals radii (McCammon *et al.*, 1979) and a probe radius of 1.4 Å unless specified otherwise. This method has the advantage that if a cavity is present within a protein or was created by the removal of a cofactor from the coordinate file (see above), the volume of the cavity is not included as part of the overall protein volume (*i.e.* the Connolly algorithm gives the volume of only the remaining proteinaceous material). In cases in which the functional molecule comprised multiple asymmetric units (1rop, 1rsc, 1stm, 2fua, 2gyi and 3fua), volume calculations were restricted to a single asymmetric unit.

In the second method for calculating molecular volumes, Voronoi volumes for all atoms in the protein were determined with the program *calcvolume.exe* (Harpaz *et al.*, 1994; Gerstein *et al.*, 1995) using the normal Voronoi algorithm (method 1). Atoms for which closed polyhedra could not be formed were omitted from further consideration. The remaining atoms were then identified as either 'exposed' or 'buried' and the corresponding atomic volumes were summed for each category. To determine if a given atom was on the protein surface, the solvent-accessible area was calculated using the program *calcsurface.exe* (Gerstein, 1992) using a probe radius of 1.4 Å. 'Exposed' atoms were defined as those with a solvent-accessible area greater than zero. All other atoms were considered 'buried'.

Molecular weights were calculated using *MWC* (M. L. Quillin, unpublished program). Contributions from H atoms were added implicitly (*i.e.* if a non-H atom was included in the calculation of molecular weight, then all H atoms directly bonded to the given atom were also included). For the purposes of these calculations, arginine, lysine and

histidine residues and N-termini were assumed to be protonated, while aspartate and glutamate residues and C-termini were assumed to be unprotonated. Cysteine residues were assumed to be in the reduced form, although this is clearly not the case in some structures. Because of the low atomic mass of hydrogen, however, the errors in molecular weight attributable to these assumptions are expected to be small (<0.1%).

Protein densities were calculated according to the formula $\rho = 10^{24}M/(VN_A)$, where ρ is the protein density in g cm^{-3}, M is the molecular weight in g mol^{-1}, V is the molecular volume in Å^3 and N_A is Avogadro's number (6.022×10^{23} mol^{-1}). Atoms that were not included in volume calculations were also omitted from molecular-weight calculations.

A possible source of systematic error in density calculations based on volumes obtained with the Connolly algorithm concerns the choice of probe radius. It appears, however, that density values calculated using this method are relatively insensitive to the exact value used for the probe radius, decreasing monotonically with probe radius such that a 0.1 Å increase in probe radius produces an approximately 0.02 g cm^{-3} decrease in density (data not shown).

3. Results and discussion

The value for the mean partial specific volume of proteins determined from solution studies is 0.73 cm^3 g^{-1}, which corresponds to a density of 1.37 g cm^{-3} (Gekko & Noguchi, 1979; Kuntz & Kauzmann, 1974; Squire & Himmel, 1979). Using the algorithm of Connolly (1993) to calculate molecular volumes, the mean density for the 30 representative proteins shown in Table 1 was found to be 1.43 ± 0.03 g cm^{-3}, which is in reasonably good agreement with the experimental value. In contrast, using a procedure based on the Voronoi algorithm, Andersson & Hovmöller (1998) obtained a much lower value for the same set of proteins, namely 1.22 ± 0.02 g cm^{-3}.

To determine the source of this discrepancy, we carried out independent density calculations with the same implementation of the Voronoi algorithm (Harpaz *et al.*, 1994; Gerstein *et al.*, 1995) used by Andersson & Hovmöller, but with special attention to atoms that are at or near the surface (see §2). As shown in Table 1, by limiting density calculations to atoms that are strictly buried, a mean protein density of 1.47 ± 0.05 g cm^{-3} is obtained. This is much higher than the value of 1.22 ± 0.02 g cm^{-3} obtained by Andersson & Hovmöller (1998).

We suggest that the disparity between the value for the mean density of proteins calculated by Andersson & Hovmöller and that obtained in the present study is a consequence of a well known caveat of the Voronoi algorithm. In order to define a Voronoi polyhedron for a given atom, the atom in question must be surrounded by other atoms. For atoms close to or at the surface, however, one or more vertices of the polyhedron can sometimes extend into bulk solvent, giving a spuriously high estimate of the atomic volume. The substantial difference between the Voronoi volumes calculated for atoms that are solvent exposed and for atoms that are fully buried is shown in Fig. 1.

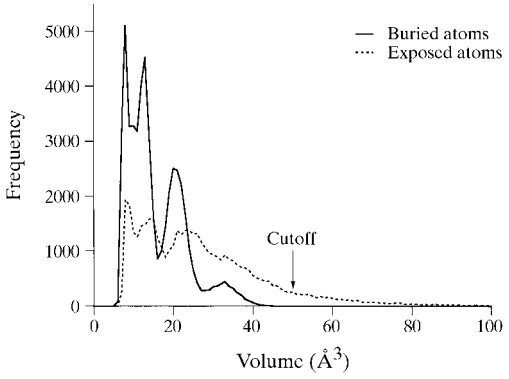

Figure 1
Histograms of Voronoi volumes of buried and exposed atoms, where 'buried' and 'exposed' are as defined in §2. The solid line represents the distribution of volumes for atoms strictly buried within proteins. The dashed line corresponds to atoms that have some accessibility to solvent but still permit the calculation of a Voronoi volume. The arrow shows the 50 Å^3 cutoff used by Andersson & Hovmöller (1998).

Andersson & Hovmöller (1998) were aware of the above limitation of the Voronoi procedure and attempted to deal with it by rejecting atoms for which the calculated volume exceeded 50 Å^3 (2–3% of the total number of atoms according to Andersson & Hovmöller, 1998). As is clear from Fig. 1, however, the 50 Å^3 cutoff still admits many surface atoms that have inflated volumes, leading in turn to an erroneously low value for the calculated density of the protein. In order to emphasize the magnitude of the volume inflation, we have included in Table 1 the densities calculated for the atoms within each protein that are exposed to solvent, but are still sufficiently surrounded by other atoms to permit the calculation of a Voronoi volume. In this case, the mean density is 0.87 ± 0.04 g cm^{-3}, unquestionably an underestimate of the true value.

That the Voronoi procedure itself is not flawed is also shown by the results of Harpaz *et al.* (1994) (see also Tsai *et al.*, 1999). These authors determined the mean values for the Voronoi volumes of buried residues and used them to calculated an average partial specific volume of 0.728 cm^3 g^{-1} for 13 representative proteins. This corresponds to a density of 1.37 g cm^{-3}, which is in excellent agreement with the average experimental value.

Atomic volumes and derived protein densities are of fundamental importance in the analysis of protein structure. In addition, many crystallographic procedures rely upon accurate values for these quantities either implicitly or explicitly. For these reasons, it is critical that the systematic error associated with Voronoi volume calculations for surface atoms be recognized and treated in a rational, consistent manner. Moreover, since we have shown that there is no need to revise the currently accepted value for the average density of proteins, there is also no basis to revise (Andersson & Hovmöller, 2000) the relationship (Matthews, 1968) between the parameter V_M (Matthews, 1968, 1974) and the solvent content of protein crystals.

We thank Doug Juers, Richard Kingston and Mark Gerstein for useful discussions. MLQ acknowledges the support of a fellowship from the Helen Hay Whitney Foundation. This work was also supported in part by NIH grant GM20066 to BWM.

References

Andersson, K. M. & Hovmöller, S. (1998). *Z. Kristallogr.* **213**, 369–373.
Andersson, K. M. & Hovmöller, S. (2000). *Acta Cryst.* D**56**, 789–790.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Connolly, M. L. (1993). *J. Mol. Graph.* **11**, 139–141.
Finney, J. L. (1975). *J. Mol. Biol.* **96**, 721–732.
Gekko, K. & Noguchi, H. (1979). *J. Phys. Chem.* **83**, 2706–2714.
Gerstein, M. (1992). *Acta Cryst.* A**48**, 271–276.
Gerstein, M. & Chothia, C. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 10167–10172.
Gerstein, M., Tsai, J. & Levitt, M. (1995). *J. Mol. Biol.* **249**, 955–966.
Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Structure*, **2**, 641–649.
Kauzmann, W., Moore, K. & Schultz, D. (1974). *Nature (London)*, **248**, 447–449.

Kuntz, I. D. Jr & Kauzmann, W. (1974). *Adv. Protein Chem.* **28**, 239–345.
Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.
McCammon, J. A., Wolynes, P. G. & Karplus, M. (1979). *Biochemistry*, **18**, 927–942.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Matthews, B. W. (1974). *Annu. Rev. Phys. Chem.* **27**, 493–523.
Pontius, J., Richelle, J. & Wodak, S. J. (1996). *J. Mol. Biol.* **264**, 121–136.
Richards, F. M. (1974). *J. Mol. Biol.* **82**, 1–14.
Squire, P. G. & Himmel, M. E. (1979). *Arch. Biochem. Biophys.* **196**, 165–177.
Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). *J. Mol. Biol.* **290**, 253–266.
Voronoi, G. F. (1905). *J. Reine Angew. Math.* **134**, 198–287.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.