J. Mol. Biol. (1996) **264**, 121–136

JMB

Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures

Joan Pontius, Jean Richelle and Shoshana J. Wodak*

Unité de Conformation de Macromolécules Biologiques Université Libre de Bruxelles Av. F. Roosevelt 50 CP160/16, B-1050 Bruxelles Belgium

Standard ranges of atomic and residue volumes are computed in 64 highly resolved and well-refined protein crystal structures using the classical Voronoi procedure. Deviations of the atomic volumes from the standard values, evaluated as the volume *Z*-scores, are used to assess the quality of protein crystal structures. To score a structure globally, we compute the volume *Z*-score root mean square deviation (*Z*-score rms), which measures the average magnitude of the volume irregularities in the structure. We find that the *Z*-score rms decreases as the resolution and *R*-factor improve, consistent with the fact that these improvements generally reflect more accurate models. From the *Z*-score rms distribution in structures with a given resolution or *R*-factor, we determine the normal limits in *Z*-score rms values for structures solved at that resolution or *R*-factor. Structures whose *Z*-score rms exceeds these limits are considered as outliers. Such structures also exhibit unusual stereochemistry, as revealed by other analyses. Absolute *Z*-scores of individual atoms are used to identify problems in specific regions within a protein model. These *Z*-scores correlate fairly well with the atomic *B*-factors, and atoms having absolute *Z*-scores >3, occur at or near regions in the model where programs such as PROCHECK identify unusual stereochemistry. Atomic volumes, themselves not directly restrained in crystallographic refinement, can thus provide an independent, rather sensitive, measure of the quality of a protein structure. The volume-based structure validation procedures are implemented in the program PROVE (PROtein Volume Evaluation), which is accessible through the World Wide Web.

© 1996 Academic Press Limited

**Corresponding author*

Keywords: Voronoi volumes; protein structures; quality assessment

Introduction

Assessing the quality of the 3D structure of a macromolecule determined by methods such as X-ray diffraction and nuclear magnetic resonance techniques has become an important issue, as the information compiled from the rapidly growing number of solved structures is increasingly used in many applications ranging from drug design to protein structure predictions.

Several procedures have been investigated recently for use in detecting inaccuracies in protein 3D structures (MacArthur *et al*., 1994). Most of these work by compiling statistics of selected parameters in high quality protein structures. Test structures are then analyzed by measuring how similar their parameters are to the standard values derived from the high quality ones. However, refinement programs often use the same parameters as constraints or restraints. Least squares refinement algorithms usually use restraints on covalent geometry (PROLSQ, TNT) (Hendrickson & Konnert, 1980; Tronrud *et al*., 1987), or constraints and restraints on groups of atoms (CORELS) (Sussman, 1985). Restraints on non-bonded contacts are in addition, applied in procedures based on molecular-mechanics (EREF) (Jack & Levitt, 1978), and molecular-dynamics (XPLOR) (Brünger *et al*., 1987) algorithms. The standard values for the restrained parameters are derived either from the analysis of small molecules (Engh & Huber, 1991) or from molecular mechanics and molecular dynamics force-fields (Brooks *et al*., 1983).

These restraints and constraints can leave their mark on the final model. It has, for example, been

Abbreviations used: PROVE, PROtein Volume Evaluation; PLL, oncogene protein; GNS, gene SDNA binding protein; PGM, phosphoglycerate mutase; CY3, cytochrome *c*3; ABX, alpha-bungarotoxin; LDH, lactate dehydrogenase; 3D, three-dimensional; vdW, van der Waals; PDB, Protein Data Bank.

0022–2836/96/460121–16 $25.00/0

© 1996 Academic Press Limited

shown that the frequency of *cis*-prolines increases in higher resolution protein structures (Stewart *et al*., 1990), which implies that the refinement programs which restrain proline to the *trans* position need to be re-evaluated. Measuring the quality of a structure in terms of how well certain parameters match the standard values may have pitfalls, as it may actually evaluate how different standard values compare with one another (Laskowski *et al*., 1993a,b).

These issues suggest the need for objective methods for assessing the quality of a protein model. Such assessment should be able to assign an objective reliability score for a structure as a whole, as well as detect regions of irregularities within the structure. It should therefore be based on parameters that are not directly restrained during refinement. When a parameter, not restrained during refinement, has its value changing as a function of the crystallographic resolution, or, if it can detect misfolded proteins, it becomes an interesting candidate to use as a structure quality measure.

Procedures using variables not directly restrained during refinement have included analysis of residue surface area, polarity and secondary structure (Luthy *et al*., 1992), atomic solvation (Holm & Sander, 1992), and distances between C^{β} atoms (Sippl & Weitckus, 1992). By analyzing over 400 protein structures, Morris *et al*. (1992) and Laskowski *et al*. (1993a,b) were able to define stereochemical parameters that change as the resolution improves. Among these were the percentage of residues found in defined regions of the Ramachandran plot, close contacts between residues, and hydrogen bond energies. These observations are now consolidated in the package PROCHECK, which assesses the quality of a given protein structure by comparing its stereochemical parameters with those in structures judged to be acceptable standards.

Here, we show that atomic volumes can be added to the list of independent parameters which can be used to measure the quality of a 3D structure. Atomic volumes are clearly influenced by a number of parameters (bond distances, bond angles, non-bonded contacts) that are subject to restraints in many refinement procedures. However, volumes as such, are not restrained during refinement. Normal ranges of atomic volumes (Finney, 1975; Gellatly & Finney, 1982) and residues (Chothia, 1975; Gerstein *et al*., 1994; Harpaz *et al*., 1994) have been previously computed, but have not been regularly updated, and not used in the context of structure quality assessment.

The volumes of atoms and residues in proteins were first computed by Richards (1974) and then Finney (1975), using procedures based on the Voronoi method (Voronoi, 1908). This method uses atomic positions derived from X-ray diffraction experiments, and the volume assigned to each atom is defined as the smallest polyhedron created by the set of planes that bisect the vectors connecting the center of the atom to those of its neighbors. Its first (Richards, 1974) and more recent (Gellatly & Finney, 1982) applications to proteins often involved modifications, which take into account the difference in vdW radii between atoms. This requires assigning a consistent set of radii to atoms in proteins, which is a difficult problem because proteins are a highly heterogeneous medium. Furthermore there is little hope for obtaining an adequate set of atomic radii for all the ligands and co-factors encountered in protein crystal structures.

In this work we therefore use the parameter-free classical Voronoi procedure, as implemented in SurVol (Alard, 1991), to compute the normal ranges for atomic and residue volumes in a set of 64 highly resolved and well-refined protein structures. We then investigate the use of these normal ranges in assessing the quality of a protein structure by

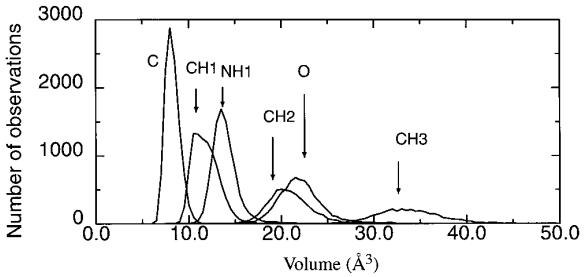
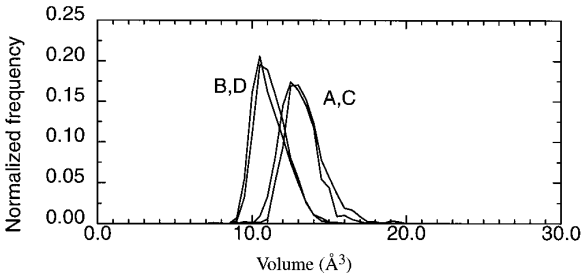
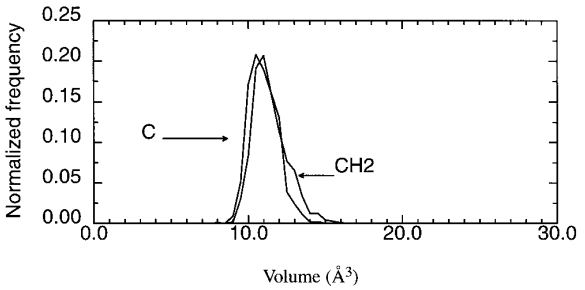


Figure 1. Volume distributions for six different subpopulations of atoms. Distributions of atomic volumes computed with the Voronoi procedure for atom subpopulations segregated according to their chemical types. We use the 23 chemical types defined in the modelling package BRUGEL (Delhaise *et al*., 1985). They are as follows: C: sp2 carbon of the peptide, amide and carboxylate groups; CH1: sp3 carbon connected to one hydrogen (C^{α} of most amino acids); CH2: sp3 carbon connected to two hydrogens (C^{β} of most amino acids); CH3: sp3 carbon connected to three hydrogens (e.g. $C^{\delta 1}$ and $C^{\delta 2}$ of Leu); CR15: sp2 carbon connected to one hydrogen in 5-atom rings ($C^{\epsilon 1}$ and $C^{\delta 2}$ of His, $C^{\delta 1}$ of Trp); CR16: sp2 carbon connected to one hydrogen in six-atom rings (e.g. $C^{\delta 1}$, $C^{\delta 2}$, $C^{\epsilon 1}$ and $C^{\epsilon 2}$ of Phe); CR5: sp2 carbon without hydrogen in five-atom rings (C^{γ} of His and C^{γ} of Trp); CR56: sp2 carbon between two rings, one of five atoms and one of six atoms ($C^{\delta 2}$ and $C^{\epsilon 2}$ of Trp); CR6: sp2 carbon without hydrogen in six atom rings (C^{γ} Phe and Tyr, C^{ζ} of Tyr); N: sp2 nitrogen without hydrogen (main-chain N of Pro); NC1: sp2 nitrogen connected to one hydrogen in a charged group (N^{ϵ} of Arg); NC2: sp2 nitrogen connected to two hydrogens in a charged group ($N^{\eta 1}$, $N^{\eta 2}$ of Arg); NC3: sp3 nitrogen connected to three hydrogens in a charged group (e.g. N^{ζ} of Lys, amino terminus); NH1: sp2 nitrogen connected to one hydrogen (main-chain N); NH2: sp2 nitrogen connected to two hydrogens ($N^{\delta 2}$ of Asn, $N^{\epsilon 2}$ of Gln); NRD5: sp2 nitrogen without hydrogen in five-atom rings ($N^{\epsilon 2}$ of His and $N^{\epsilon 1}$ of Trp); NR15: sp2 nitrogen connected to one hydrogen in five-atom rings ($N^{\epsilon 1}$ of His, $N^{\epsilon 1}$ of Trp); O: sp2 oxygen one without net charge (main-chain O); OC: sp2 oxygen with a net charge ($O^{\epsilon 1}$ and $O^{\epsilon 2}$ of Glu, $O^{\delta 1}$ and $O^{\delta 2}$ of Asp, and chain ends); OH1: oxygen of alcohol groups in side-chains; OH2: oxygen of water; S: sulfur without hydrogen; SH1: sulfur with hydrogen.



(a)



(b)

Figure 2. Influence of the nature of the covalently bonded neighbors on the atomic volumes computed with the classical Voronoi procedure. (a) Four volume distributions corresponding to atoms of chemical type CH1 bonded to atoms of different chemical types. A, Atoms bonded to atoms of chemical type CH1, CH3 and CH3 (Val C^{β} atoms). B, Atoms bonded to atoms of chemical type C, N, CH1 (C^{α} atoms with CH1 in the β position: Val, Ile and Thr). C, Bonded to atoms of chemical type CH2, CH3, CH3 (Leu C^{γ} atoms). D, Bonded to C, N and CH2 (C^{α} atoms of all the residues with a CH2 in the β position except Pro). (b) Volumes of C^{α} atoms with CH2 in β position. Distributions calculated according to the chemical type of the atom in the γ position. Shown are the distribution of C^{α} when the C^{γ} atom is C (Asn or Asp) and CH_2 (Lys, Gln, Arg, Glu, Met).

comparing the volumes of its atoms to the corresponding standard values.

Given the heterogeneous nature of the protein medium, departures from standard values can be expected to occur for physically meaningful reasons, and thus not be necessarily due to model inaccuracies. Factors affecting atomic volumes are therefore investigated in an effort to focus on volume irregularities that reflect the quality of a reported structure as a whole, and of regions within a given structure. Using a test set of 900 protein structures, spanning an appreciable range of resolutions (1 to 3.9 Å) and *R*-factors (0.098 to 0.430), we then show how the structure quality assessment based on these irregularities, which are implemented in the program PROVE (PROtein Volume Evaluation), compare with, and complement, other commonly used structure validation criteria, such as those provided by PROCHECK.

Results

Factors influencing the volumes of buried protein atoms

Atomic volumes were computed using the classical Voronoi method, for all the buried atoms in the 64 protein structures of our high resolution reference set, as described in Methods.

With the aim of deriving standard volumes, the computed atomic volumes were analyzed for the presence of subpopulations, and the parameters influencing the distributions in these subpopulations were investigated.

A first useful influence to analyze was that of the atom chemical type, which reflects its bonding properties and chemical character. Here we consider the 23 atom chemical types used in the BRUGEL package (Delhaise *et al*., 1985), which are listed in the legend to Figure 1. It is not surprising to see that the computed volumes show distinct subpopulations which depend on the atom chemical type (Figure 1). We see, in particular, that the volume of atoms correlate better with their bonding properties than with their vdW radii. For example, the CH3 group, which is usually considered as having the same vdW radius as the CH1 and CH2 groups (Richards, 1974; Finney, 1975; Chothia, 1975; Brooks *et al*., 1983), but is bonded to only one atom, has a larger volume than all other groups. Similarly, the backbone carbonyl oxygen, being bonded to only one atom, has a larger volume than the backbone amide (NH1), even though it is considered to have a smaller vdW radius than other backbone atoms.

The influence of the number and nature of the covalently bonded neighbors on the computed atomic volume is further illustrated in Figure 2. Figure 2(a) shows four volume distributions (A to D), each corresponding to an atom with the CH1 chemical-type, in different bonding environments. A and C correspond to the Val C^{β} and Leu C^{γ}, respectively, and B and D, to the C^{α} in Val, Ile, Thr, and the C^{α} in all other residues except Pro.

We see that the B and D distributions for the C^{α} volumes, on the one hand, and the A and C distributions for the Val C^{β} and Leu C^{γ} volumes on the other, virtually superimpose, but that the two pairs of distributions are quite distinct. The former two distributions have a mean volume of $\sim$11.5 Å^3 and the latter, a mean volume of $\sim$13.3 Å^3, representing a difference in mean volume of $\sim$14%. An analysis of C^{α} atoms having CH_2 in the C^{β} position showed that there is very little influence on the volume from the chemical type of atoms two bonds away. The greatest difference between C^{α} distributions based on the chemical type of the atom two bonds away is between those having a C and a CH2 in the γ position. However, as illustrated in Figure 2(b), this difference is negligible, with the

Table 1. Standard volumes of atoms in proteins

Atom type		⟨V⟩ (σ)	N	Atom type		⟨V⟩ (σ)	N
Ala	N_NH1	14.4(1.5)	871	Gly	O_O	22.4(2.6)	553
	CA_CH1	12.4(1.3)	772		O_OC	26.8(0.0)	1*
	C_C	8.6(0.7)	983	His	N_NH1	13.8(1.4)	307
	O_O	22.8(2.4)	631		CA_CH1	11.4(1.0)	225
	CB_CH3	32.7(3.0)	425		C_C	8.4(0.7)	322
	O_OC	18.1(4.8)	2*		O_O	21.6(2.5)	226
Arg	N_NH1	14.0(1.3)	410		CB_CH2	20.5(2.1)	189
	CA_CH1	11.7(1.2)	302		CG_CR5	10.0(0.7)	224
	C_C	8.4(0.7)	420		ND1_NR15	16.5(2.2)	156
	O_O	22.2(2.2)	265		CD2_CR15	19.7(2.6)	168
	CB_CH2	20.4(1.8)	200		CE1_CR15	19.2(2.5)	86
	CG_CH2	20.9(2.3)	191		NE2_NRD5	17.8(2.8)	86
	CD_CH2	20.3(2.7)	131	Ile	N_NH1	14.1(1.1)	714
	NE_NC1	16.3(1.7)	129		CA_CH1	11.4(1.1)	627
	CZ_C	9.4(0.8)	163		C_C	8.2(0.6)	749
	NH1_NC2	23.1(3.0)	78		O_O	22.5(2.1)	542
	NH2_NC2	24.1(3.2)	59		CB_CH1	13.0(1.0)	634
Asn	N_NH1	13.8(1.4)	543		CG1_CH2	22.4(2.0)	521
	CA_CH1	11.2(0.9)	303		CG2_CH3	33.2(2.9)	402
	C_C	8.5(0.7)	477		CD1_CH3	35.3(3.4)	374
	O_O	22.0(2.4)	273		N_NC3	22.9(0.0)	1*
	CB_CH2	20.2(2.2)	158	Leu	N_NH1	14.1(1.3)	1095
	CG_C	9.2(0.7)	186		CA_CH1	11.6(0.9)	943
	OD1_O	21.7(2.9)	143		C_C	8.5(0.7)	1074
	ND2_NH2	25.5(4.6)	86		O_O	22.3(2.3)	742
Asp	N_NH1	13.9(1.5)	602		CB_CH2	20.8(1.6)	826
	CA_CH1	11.3(0.9)	303		CG_CH1	13.7(1.3)	977
	C_C	8.3(0.6)	629		CD1_CH3	35.2(3.3)	611
	O_O	22.2(2.8)	349		CD2_CH3	35.0(3.5)	545
	CB_CH2	20.8(2.0)	206	Lys	N_NH1	14.0(1.3)	659
	CG_C	9.1(0.8)	182		CA_CH1	11.6(1.0)	369
	OD1_OC	20.7(3.1)	165		C_C	8.5(0.7)	655
	OD2_OC	21.6(3.7)	133		O_O	22.3(2.4)	362
Cysh	N_NH1	14.3(1.4)	101		CB_CH2	20.6(1.9)	227
	CA_CH1	11.8(1.3)	86		CG_CH2	21.0(2.1)	180
	C_C	8.5(0.6)	107		CD_CH2	21.7(2.3)	117
	O_O	22.8(2.6)	71		CE_CH2	21.8(2.5)	49
	CB_CH2	22.8(2.1)	65		NZ_NC3	23.0(4.1)	16*
	SG_SH1	34.2(3.7)	54	Met	N_NH1	14.0(1.5)	230
Cyss	N_NHI	14.3(1.7)	114		CA_CH1	11.7(1.1)	188
	CA_CH1	11.4(0.8)	119		C_C	8.5(0.7)	241
	C_C	8.4(0.6)	130		O_O	22.6(2.5)	182
	O_O	21.7(2.0)	78		CB_CH2	21.0(1.9)	175
	CB_CH2	21.2(2.2)	90		CG_CH2	23.0(1.9)	166
	SG_S	25.0(2.6)	97		SD_S	27.8(2.7)	159
Gln	N_NHI	13.9(1.4)	337		CE_CH3	34.8(4.4)	114
	CA_CH1	11.5(1.0)	190		N_NC3	37.4(0.7)	2*
	C_C	8.4(0.7)	325	Phe	N_NH1	14.0(1.3)	476
	O_O	22.0(2.4)	204		CA_CH1	11.6(1.1)	380
	CB_CH2	20.0(1.8)	157		C_C	8.4(0.8)	486
	CG_CH2	20.4(2.2)	121		O_O	22.5(2.1)	353
	CD_C	9.6(0.7)	101		CB_CH2	21.1(1.9)	361
	OE1_O	23.4(3.2)	72		CG_CR6	10.2(0.7)	531
	NE2_NH2	24.6(3.8)	71		CD1_CR16	20.3(2.2)	387
Glu	N_NH1	14.0(1.3)	487		CD2_CR16	20.8(2.0)	371
	CA_CH1	11.7(1.2)	283		CE1_CR16	22.0(2.3)	337
	C_C	8.4(0.6)	540		CE2_CR16	22.2(2.2)	307
	O_O	22.1(2.4)	249		CZ_CR16	22.0(2.4)	308
	CB_CH2	20.3(1.8)	172		O_OC	19.3(0.0)	1*
	CG_CH2	21.1(2.6)	124	Pro	N_N	9.4(0.7)	523
	CD_C	9.1(0.7)	82		CA_CH1	12.1(1.1)	310
	OE1_OC	22.8(3.3)	49		C_C	8.4(0.6)	412
	OE2_OC	23.5(3.8)	53		O_O	22.8(2.9)	232
Gly	N_NH1	14.9(1.6)	657		CB_CH2	23.0(2.3)	151
	CA_CH2	20.0(2.0)	348		CG_CH2	24.1(2.9)	124
	C_C	9.3(0.8)	598		CD_CH2	20.7(2.3)	162

continued

Table 1. *continued*

Atom type		$\langle V \rangle$ (σ)	*N*	Atom type		$\langle V \rangle$ (σ)	*N*
Ser	N_NH1	14.3(1.4)	585	Tyr	N_NH1	13.9(1.2)	446
	CA_CH1	11.6(1.2)	425		CA_CH1	11.5(1.0)	380
	C_C	8.4(0.7)	639		C_C	8.5(0.8)	464
	O_O	22.1(2.6)	438		O_O	22.1(2.2)	334
	CB_CH2	20.9(2.2)	246		CB_CH2	21.3(2.0)	325
	OG_OH1	23.4(3.4)	213		CG_CR6	10.2(0.7)	483
Thr	N_NH1	14.0(1.2)	614		CD1_CR16	20.0(2.0)	348
	CA_CH1	11.3(1.1)	508		CD2_CR16	20.1(2.1)	317
	C_C	8.3(0.6)	687		CE1_CR16	20.3(2.1)	237
	O_O	21.9(2.3)	413		CE2_CR16	20.3(2.0)	215
	CB_CH1	12.9(1.2)	245		CZ_CR6	10.0(0.7)	281
	OG1_OH1	23.1(3.3)	203		OH_OH1	25.1(4.1)	116
	CG2_CH3	31.8(2.9)	169	Val	N_NH1	14.1(1.2)	1029
Trp	N_NH1	14.3(1.4)	168		CA_CH1	11.4(1.1)	947
	CA_CH1	11.8(1.0)	947		C_C	8.4(0.7)	1108
	C_C	8.4(0.9)	185		O_O	22.6(1.9)	724
	O_O	22.1(2.4)	135		CB_CH2	13.3(1.1)	904
	CB_CH2	21.5(2.0)	132		CG1_CH3	33.5(2.9)	572
	CG_CR5	10.4(0.6)	181		CG2_CH3	33.4(2.8)	644
	CD1_CR15	20.1(2.2)	108				
	CD2_CR56	10.8(0.7)	181				
	NE1_NR15	18.3(2.5)	87				
	CE2_CR56	10.2(0.8)	165				
	CE3_CR16	20.8(1.9)	153				
	CZ2_CR16	20.9(2.1)	90				
	CZ3_CR16	22.1(2.2)	132				
	CH2_CR16	21.4(2.1)	111				

Listed are the average values and the standard deviations of atomic volumes computed using the classical Voronoi procedure, applied to the reference set of 64 highly resolved and well-refined protein structures (Table 4), as described in Methods. Only completely buried atoms were considered in the calculations. The leftmost column gives the amino acid name. The second column lists the atom names with the IUPAC code (left) separated by an underscore from the chemical type code (right). O_OC and N_NC3 atom types, in Ala, Gly, Ile, Met, Phe, denote terminal backbone groups. Column 4 gives the average volumes $\langle V \rangle$ and standard deviation σ (in parentheses). The rightmost column gives the number of observations used in the calculations. Atom types with less than 20 observations were not used in *Z*-score calculations. These atom types refer to terminal groups, and other charged atoms which are rarely buried.

means of the corresponding distributions being 11.2 Å^3 and 11.6 Å^3, respectively.

Standard volumes of atoms and residues

The mean volumes of buried atoms and their standard deviations were computed from our set of 64 highly resolved and refined proteins, using the classical Voronoi procedure and considering subpopulations corresponding to identical atom types. This segregation considers, for example, as distinct populations, the backbone oxygens of different residues, and distinguishes between the backbone carbonyl oxygens, and those at C termini. These data, which represent the standards against which atomic volumes of test proteins are compared, are given in Table 1.

Residue volumes were computed by summing the volumes of their component atoms. The calculations took into account only completely buried residues, representing a small fraction of the total number (∼6%). Table 2 lists the mean residue volumes computed in this study using the Voronoi method alongside residue volumes computed by other authors. The older values (Chothia, 1975) were computed on a smaller protein sample, considering atoms which had between 0 and 5% of their surface exposed to solvent. The more recent values of Harpaz *et al*. (1994) were computed from a set of 108 highly resolved protein structures, considering, as here, only buried atoms. In these two previous studies, the calculations were done using Richards' B method, a variant of the Voronoi method proposed by Richards (1974, 1985).

Our mean Voronoi residue volumes agree, on average, better with Chothia's values (2.9%), than with the 1994 values (3.8%). The average difference between the 1994 and 1975 values, both calculated using the same method for partitioning space, but on a different protein set and considering different subsets of atoms, is 3.7%. Also, the last row of Table 2 shows that the average variance of the residue volumes computed here with the Voronoi procedure is 4.9%, higher than that of Harpaz's residue volumes (3.7%), but lower than that of the Chothia volumes (5.9%).

Deviation from standard volumes as a measure of structure quality

Having derived the mean volumes for our different atom types, we then investigate their use

as standards with which atomic volumes computed in individual protein structures can be compared. Our aim is to use deviations from the standard volumes as a measure for the quality of a given 3D structure.

Since the resolution and the *R*-factor are good guides for the overall quality of a structure determined by X-ray diffraction, we first investigate the correlation between these parameters and the average volume irregularity of a protein structure. In addition, to examine if volume calculations can also be used to identify regions within a protein structure which are poorly modeled, we analyze the correlation between volume irregularities within a protein structure, and other measures of local structure quality such as the crystallographic *B*-factors and the stereochemical parameters used by the program PROCHECK (Morris *et al*., 1992; Laskowski *et al*., 1993b).

These various investigations are performed on structures from our test set of 900 proteins, as well as from a set of eight obsolete structures, and their replacements. Details on these sets can be found in Methods.

Deviations from standard volumes in relation to resolution

To study how the deviations from standard volumes in a protein relate to its crystallographic resolution, we compute the volume *Z*-score of each atom. This quantity is defined as the difference between the volume of the atom and the mean atomic volume for the corresponding atom type, divided by the standard deviation of the appropriate distribution. To evaluate the structure as a whole, we compute the root mean square deviation of the volume *Z*-score (*Z*-score rms) of its (buried) atoms.

Figure 3(a) displays the average *Z*-score rms computed for structures of a given resolution range as a function of resolution in the 900 protein structures of our test set. The considered resolution range was from 1 to 3.9 Å, and the averages were computed for bins of 0.1 Å resolution. We see that, for resolutions of 1.6 Å or better, the average *Z*-score rms is essentially constant, and around 1.0. Lower resolution structures display, on the average, a larger *Z*-score rms, with the average *Z*-score rms increasing steadily as the resolution decreases. The correlation factor between the average *Z*-score rms and the resolution is 0.89 over the entire range of considered resolution, and 0.98 for resolutions between 1.5 and 3.0 Å. The standard deviations of the *Z*-score rms in each resolution bin is displayed as vertical bars in Figure 3(a). They delimit the expected spread of the *Z*-score rms for a given resolution. A structure determined at a given resolution, whose *Z*-score rms falls outside the expected spread, is likely to exhibit problems.

It is noteworthy that the spread in *Z*-score rms values, for a given resolution, can be as much as 0.4,

Table 2. Residue volumes

	Voronoi		Richard's B method			
		This study		Chothia (1975)		Harpaz *et al*. (1994)
Residue	*N*		*N*		*N*	
Ala	316	91.5(5.32)	71	91.5(7.32)	387	90.1(4.66)
Arg	9	196.1(4.25)	0	NA	13	192.8(3.42)
Asn	27	138.3(5.60)	12	135.2(7.47)	41	127.5(3.29)
Asp	30	135.2(5.21)	17	124.5(6.18)	36	117.1(3.42)
Cysh	34	114.4(6.64)	4	117.7(4.16)	30	113.2(3.36)
Cyss	33	102.4(6.13)	16	105.6(5.68)	43	103.5(4.83)
Gln	12	156.4(4.32)	5	161.1(8.07)	17	149.4(3.28)
Glu	7	154.6(5.75)	13	155.1(7.35)	7	140.8(3.76)
Gly	239	67.5(5.72)	60	66.4(7.08)	323	63.8(4.55)
His	22	163.2(4.04)	8	167.3(4.42)	23	159.3(3.08)
Ile	212	162.6(3.64)	69	168.8(5.81)	234	164.9(3.76)
Leu	226	163.4(4.19)	57	167.9(6.08)	276	164.6(3.58)
Lys	4	162.5(2.24)	5	171.3(3.97)	6	170.0(3.00)
Met	56	165.9(5.31)	14	170.8(5.21)	72	167.7(4.00)
Phe	84	198.8(3.99)	29	203.4(5.06)	115	193.5(3.05)
Pro	25	123.4(6.26)	16	129.3(5.65)	64	123.1(4.79)
Ser	109	102.0(6.72)	46	99.1(7.47)	137	94.2(3.93)
Thr	66	126.0(4.92)	32	122.1(5.49)	102	120.0(4.00)
Trp	27	237.2(3.65)	9	237.6(5.72)	26	231.7(2.42)
Tyr	27	209.8(5.29)	13	203.6(4.72)	41	197.1(3.30)
Val	308	138.4(3.87)	91	141.7(5.93)	353	139.1(3.38)
Ave ($\sigma/\langle V\rangle$%)		(4.91)		(5.94)		(3.66)

Listed are the amino acid residues (column 1) and their volumes computed in three different studies. Column 3 gives the mean residue volumes computed in this study using the Voronoi procedure. Mean residue volumes computed previously using the Richards' B method by Chothia (1975) are given in column 5, and those by Harpaz *et al*. (1994) are listed in column 7. The number *N*, of residues used in computing the mean volumes, is given for each computation. The mean volumes are given in $Å^3$; values in parentheses are percentage deviations ($\sigma/\langle V\rangle$%). The bottom row gives the average percentage deviation of the residue volumes computed in each column. This shows that the Voronoi volumes computed here exhibit an average variance that is intermediate between that of the volumes of Chothia and Harpaz and colleagues.

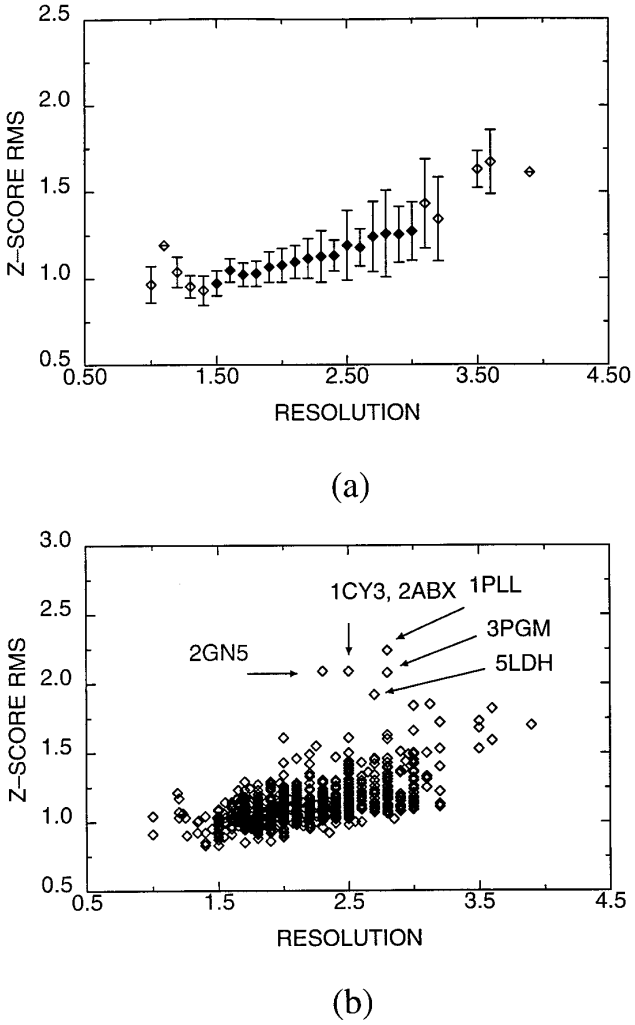


Figure 3. *Z*-score rms variation with the resolution of the crystallographic data. (a) *Z*-score rms as a function of the resolution. Average *Z*-score rms is computed for structures having the same resolution ($\pm$0.1 Å). The vertical bars indicate the magnitude of the standard deviation of the *Z*-score rms in individual resolution ranges. Graph points are derived from less than ten structures (open diamonds) and from more than ten structures (filled diamonds). (b) *Z*-score rms as in (a), displayed for individual protein structures as a function of resolution. The six furthest outlier proteins are marked by the PDB codes.

indicating that the correlation of the *Z*-score rms of individual protein structures with resolution is poorer than that of the average *Z*-score rms. This spread is clearly illustrated in Figure 3(b), which displays the *Z*-score rms of individual proteins in our test set as a function of their resolution. Comparison with Figure 3(a) shows that a number of structures have *Z*-score rms values well outside the expected spread. Six of the farthest outliers are marked (Figure 3(b)). They correspond to the entries 1PLL (oncogene protein), 2GN5 (gene 5 DNA binding protein), 3PGM (phosphoglycerate mutase), 1CY3 (cytochrome *c*3), 2ABX (alpha-bungarotoxin), and 5LDH (lactate dehydrogenase). Their *Z*-score rms values are 2.24, 2.09, 2.08, 2.09, 2.09, 1.92, respectively. All these structures were also found to be severe outliers with regard to their stereochemical parameters by PROCHECK, as summarized in Table 3.

Deviations from standard volumes in relation to the crystallographic R*-factor*

A very similar analysis, as that described above, was carried out to investigate the relation between the *Z*-score rms and the crystallographic *R*-factor. Figure 4 shows the dependence of the *Z*-score rms on the *R*-factor of the corresponding structures as quoted by their authors, together with the standard deviation of the *Z*-score rms computed for proteins having the same *R*-factor ($\pm$0.01). The correlation of the average *Z*-score rms with the *R*-factor is 0.76, markedly poorer than with the resolution. This is not unexpected, given that the *R*-factor is a versatile parameter that can be computed for various subsets of data. It hence reflects more the agreement between the model and these data subsets than the quality of the model itself. Plotting the individual *Z*-score rms as a function of the *R*-factor (data not shown) reveals the same outlier structures as in the *Z*-score rms *versus* resolution plot of Figure 3(b).

Deviations from standard volumes as an indication of problem regions in protein structures

In this section we analyze to what extent departures from standard volumes can be used to identify specific regions in a protein where the structure is of poorer quality. A first indication of whether departures from standard values of specific atoms or groups of atoms can be used as a local quality measure can be obtained by examining the correlation between the atomic volume *Z*-score and the isotropic atomic temperature factor (*B*-factor). This factor is a parameter in the crystallographic refinement, which combines many other effects in addition to thermal motion (Stroud & Fauman, 1995). Atoms with low *B*-values ($\leqslant$15) are considered as having well-defined positions, whereas those with *B*-factors $\geqslant$50 are considered as poorly defined, with intermediate *B*-factors representing intermediate levels of reliability.

Figure 5 displays the atomic volume *Z*-score rms for atoms averaged over bins of increasing *B*-factor values. We see that the average *Z*-score rms increases as the *B*-factor increases, with a correlation coefficient of 0.97, obtained ignoring the last bin ($70 < B < 80$), for which there were only 22 cases.

These results indicate that departures from standard volumes, as measured by the atomic volume *Z*-score, exhibit local variations that correlate reasonably well with those measured by the atomic *B*-factors. No such correlation could be detected between the raw atomic

Table 3. Summary of quality assessment data on outlier proteins

Structure data			This study	PROCHECK			
PDB	Resolution	*R*-factor	*Z*-score rms	% residues in allowed Ramachandran regions	H-bond energy SD	Bad contacts per 100 residue	Chi-1 pooled SD
1CY3	2.50	0.340	2.09(1.21)	25.3(76.6)	1.3(0.9)	64.4(10.5)	28.1(22.0)
2ABX	2.50	0.240	2.09(1.21)	14.8(76.6)	0.6(0.9)	41.9(10.5)	30.9(22.0)
2GN5	2.30	0.217	2.09(1.14)	45.8(79.8)	1.0(0.9)	51.7(7.6)	35.7(20.5)
3PGM	2.80	0.290	2.08(1.27)	48.5(70.9)	1.7(1.0)	63.0(15.8)	31.4(24.3)
5LDH	2.70	0.196	1.92(1.25)	56.3(72.9)	1.4(1.0)	53.2(13.9)	31.5(23.5)
1PLL	2.80	0.206	2.24(1.24)	60.3(70.9)	1.0(1.0)	2.5(15.8)	22.2(24.3)

Summary of the volume based (this study) and PROCHECK quality assessment, of the six most severe outlier proteins in Figure 3(b). The protein PDB code, resolution and *R*-factor are given in columns 1 to 3. *Z*-score rms for each structure computed in this study is given in column 4. Columns 5 to 8 list the PROCHEK evaluation: the percentage residue in allowed Ramachandran regions, the standard deviation (SD) of the H-bond energies, bad contacts per 100 residues, and the pooled standard deviation of the Chi1 side angles. For details of the PROCHECK parameters, see Morris *et al.* (1992). In parentheses are the expected values for protein structures of the same resolution.

volumes *Z*-scores, in other words, atoms with large *B*-factors showed no trend to have larger volumes.

Having established that the deviation of atomic volumes from their standard values can be used to evaluate the local quality of a protein model, we then analyzed how such evaluation compares with those based on the stereochemical parameters used in the program PROCHECK (Laskowski *et al*., 1993b). In what follows, we describe the results obtained using the Voronoi procedure for the six proteins corresponding to the furthest outliers in Figure 3(b), and for the proteins corresponding to the eight obsolete structures and their replacements.

In general, we find that atoms in residues described as outliers by PROCHECK have higher volume *Z*-scores than those not considered as outliers. For example, in obsolete structures, atoms in residues defined by PROCHECK as having (ϕ, ψ) angles outside the allowed regions of the Ramachandran map, had a *Z*-score rms of 1.40, whereas those in other residues, had *Z*-score rms of 1.21. For replacement structures, the corresponding scores were 1.07 for Ramachandran outliers, and 1.02 for those in the allowed Ramachandran regions.

Figures 6 and 7 display part of the standard PROCHECK output alongside the volume *Z*-score plots for cytochrome *c*3 (1CY3) and phosphoglycerate mutase (3PGM), two of the outliers in Figure 3(b). The *Z*-score plots show, for each residue, the largest absolute volume *Z*-score displayed by the buried atoms of this residue. This *Z*-score represents the maximum departure from the standard volume displayed in a single atom within a residue and is therefore not an average property of the residue. Figure 6c shows that residues 26, 32, 44, 47, 62, 63, and 81 of cytochrome *c*3 (1CY3) contain atoms with absolute volume *Z*-scores,

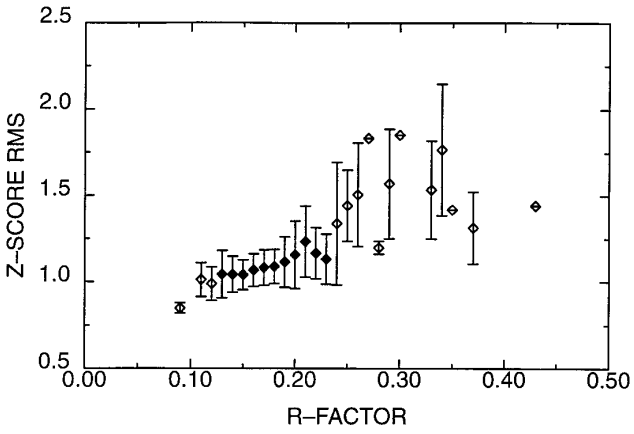


Figure 4. Variation of the *Z*-score rms as a function of the crystallographic *R*-factor, in our set of 900 protein structures. Structural *Z*-score rms values were averaged over *R*-factor bins of 0.01. Plotted are points corresponding to averages computed from less than ten observations (open diamonds) and from more than ten observations (filled diamonds). Vertical bars are standard deviations in the *Z*-score rms.

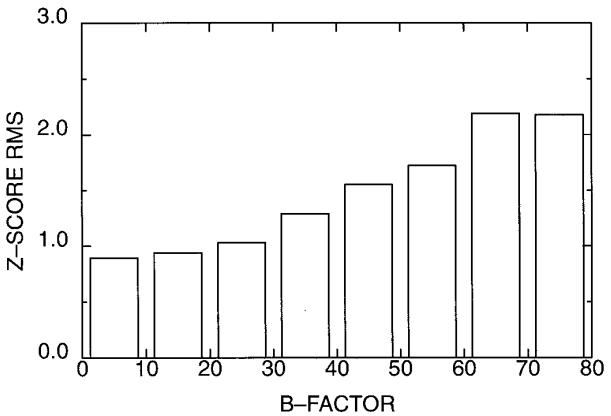


Figure 5. Behavior of *Z*-score rms of atoms as a function of *B*-factor. The absolute volume *Z*-score of atoms in the reference set of structures was averaged over *B*-factor bins of 10. *B*-factor ranges and the number of atoms in each range (in parentheses) are: 0 to 10 (15,368), 10 to 20 (29,234), 20 to 30 (7994), 30 to 40 (1690), 40 to 50 (531), 50 to 60 (221), 60 to 70 (78), 70 to 80 (22).

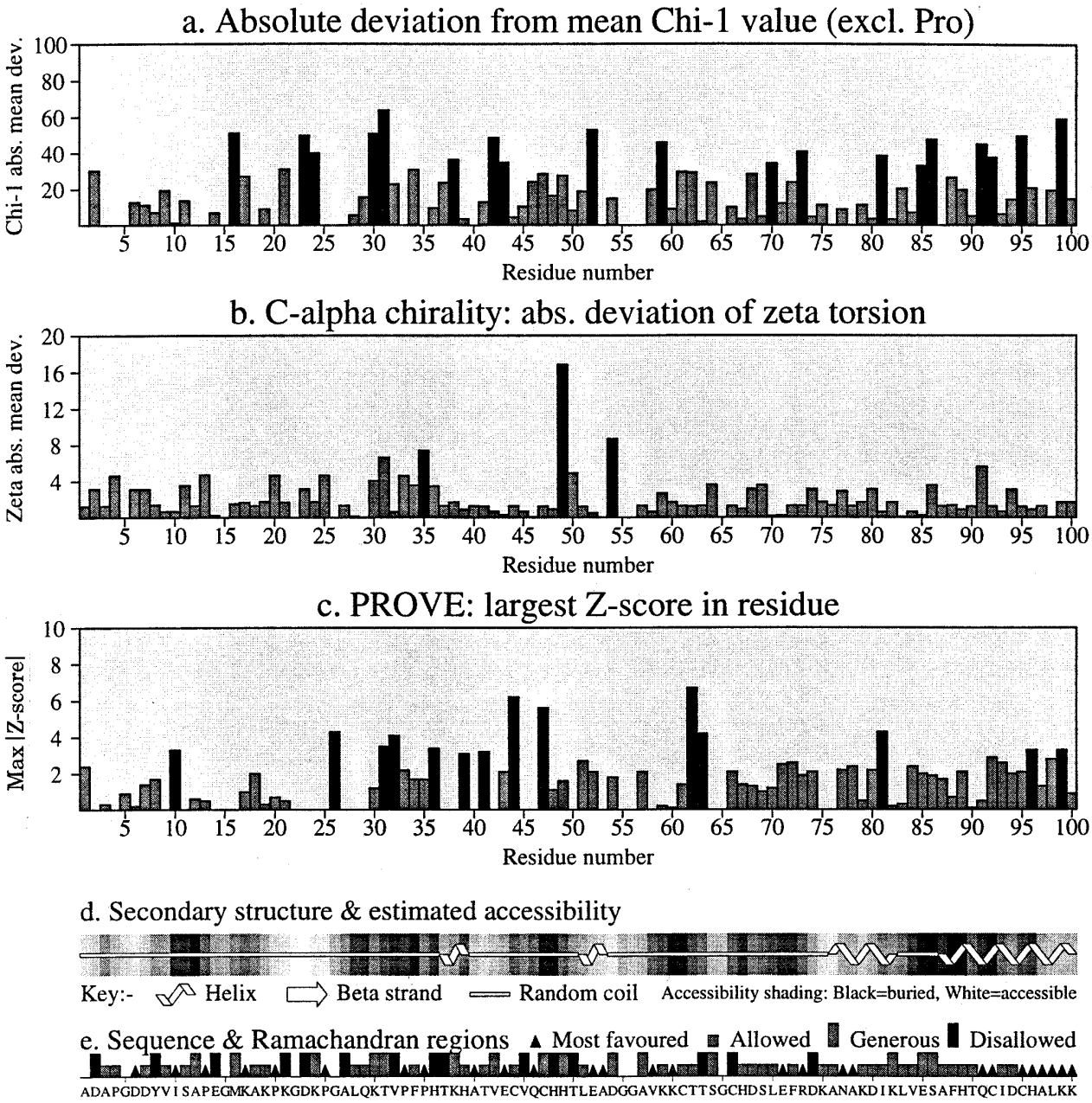


Figure 6. PROCHECK and PROVE outputs for the first 100 residues of cytochrome *c*3 (1CY3). a, Absolute deviation from mean Chi-1 value, computed by PROCHECK. Highlighted residues are those that deviate by more than two standard deviations from ideal. b, C^{α} chirality: absolute deviation of zeta torsion, computed by PROCHECK. Highlighted residues are those that deviate by more than two standard deviations from ideal. c, Maximum absolute *Z*-score of atomic volumes in individual residues along the sequence, computed by PROVE, in this study. Highlighted residues are those with *Z*-scores > 3. d, Standard PROCHECK output for secondary structure and estimated accessibility. e, Standard PROCHECK output for sequence and backbone ϕ, ψ values relative to the Ramachandran regions.

greater than 4. These high *Z*-scores belong to the backbone carbonyls of Gly26, Cys47, and Thr62, the C^{β} of Val32 and Ile81, and the backbone oxygens of Cys44, and Thr63. We find that the same residues, or their close neighbors, also have unusual Chi-1 values (residues 42 and 43 in Figure 6a), unusual omega values (residues 43 to 46, and 60 to 62, data not shown), or distorted C^{α} chirality (residue 49 and 54, in Figure 6b). Residues 45, 48, 50, 63 and 66, are also in disallowed regions of the Ramachandran map (Figure 6e). Similar observations can be made in phosphoglycerate mutase (3PGM), where residues 2, 18, 19, 20, 25, 46, 55, 59, and 61 have atoms with absolute volume *Z*-scores greater than 4 (Figure 7c). Unlike in cytochrome *c*3, these high *Z*-scores belong to both

backbone and side chain atoms. Residues 18 and 19 have unusual (ϕ, ψ) values (Figure 7e) and residues 25 to 27, 42 to 45 and 56, 58 and 59 have unusual Chi 1 values (Figure 7a).

The fact that residues found to be outliers on the basis of the volume *Z*-score of one of their atoms are not necessarily outliers by the PROCHECK measures is not surprising considering that the volume of a given atom can be affected by the position of its spatial neighbors, some of which may belong to residues far apart along the sequence. Unusual volumes may therefore result from errors occurring in several parts of the atomic model, and thereby have more complex origins than the deviations of geometric parameters such as the C^{α} chirality or the Chi 1 angle, which are due to local modelling errors.

To further illustrate the use of the atomic volume *Z*-score as a local quality measure, we analyzed the maximum atomic volume *Z*-score per residue for obsolete entries and for the corresponding replacement structures. Figure 8(a),(b) illustrate the results obtained for the first 100 residues of the obsolete and replacements entries for alcohol dehydrogenase. Figure 8(c),(d) illustrate those for the obsolete / replacement parvalbumin couple. The obsolete

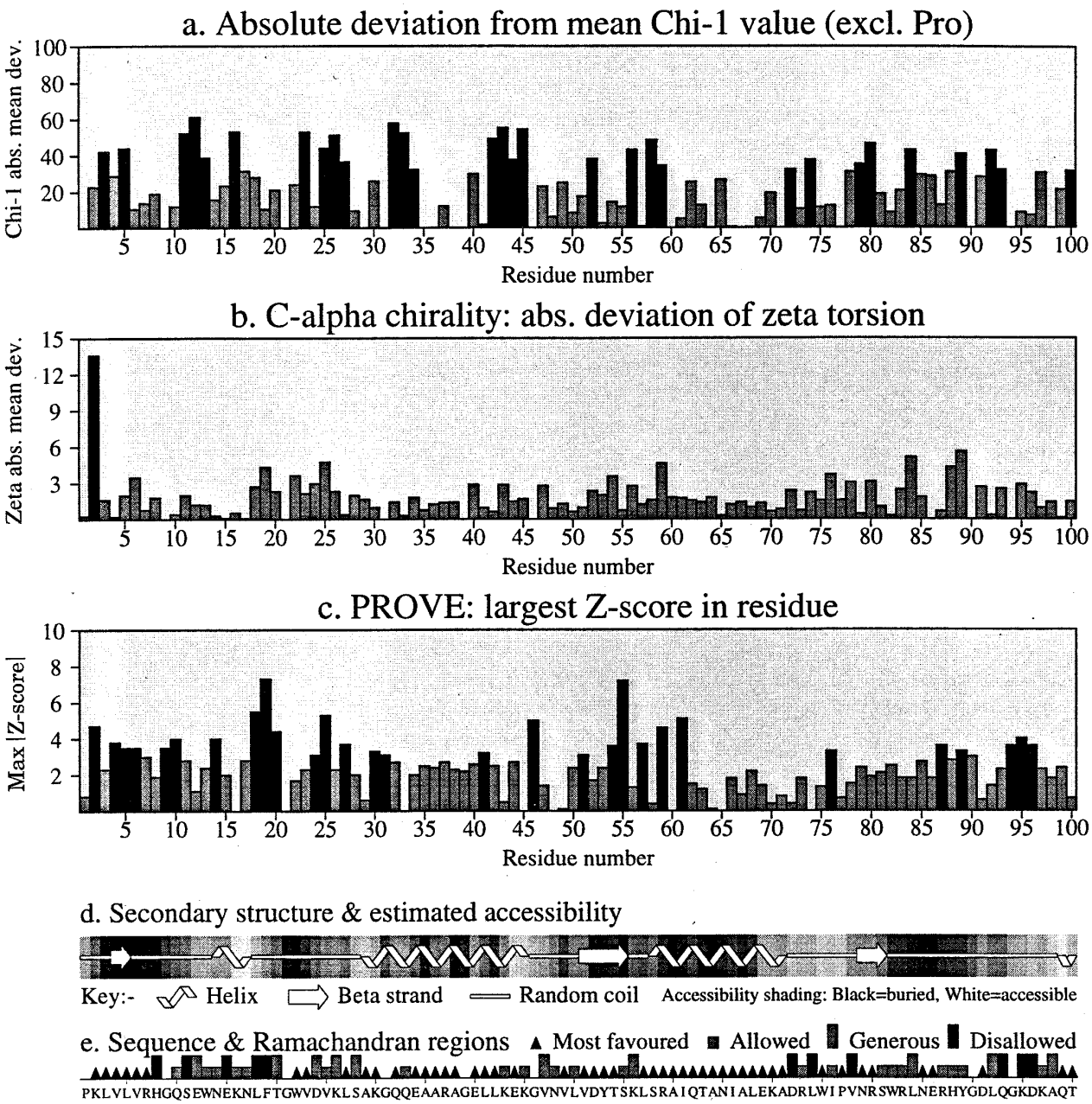


Figure 7. PROCHECK and PROVE outputs for the first 100 residues of phosphoglycerate mutase (3PGM). See legend to Figure 6 for description of plots a to e.

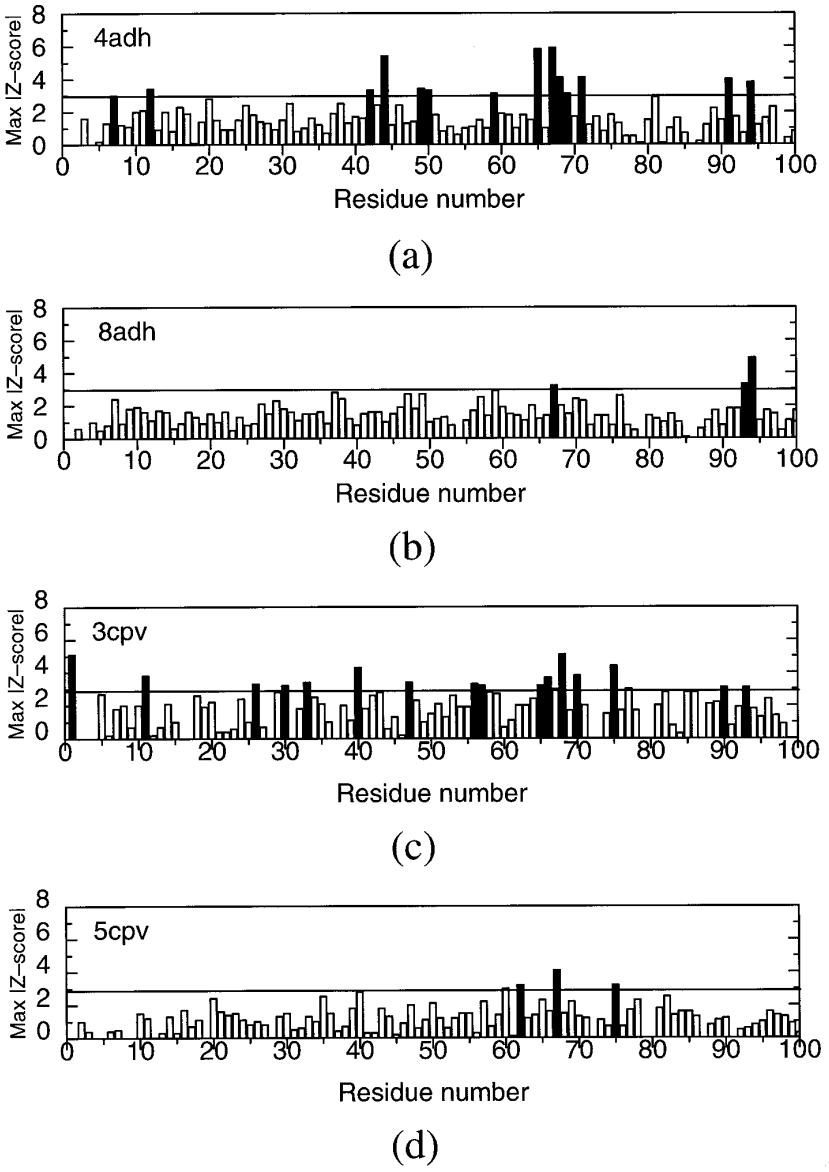


Figure 8. Atomic volume *Z*-scores in pairs of obsolete and replacement structures. Plotted are the maximum absolute *Z*-score of atomic volumes, computed by PROVE, in individual residues along the sequence for four PDB entries, corresponding to pairs of obsolete–replacement structures. Highlighted residues are those having atoms with *Z*-scores >3. (a) volume *Z*-scores for the obsolete entry for alcohol dehydrogenase (4adh); (b) volume *Z*-scores for the replacement entry (8adh); (c) volume *Z*-scores for the obsolete entry for calcium-binding parvalbumin (3cpv); (d) volume *Z*-scores for the corresponding replacement entry (5cpv). Results are given only for the first 100 residues of each structure. Details about the structures can be found in Table 5.

entry 4ADH was determined at 2.4 Å resolution, with an *R*-factor of 0.26. Its replacement (8ADH) has the same resolution, but is refined to an *R*-factor of 0.19. The *Z*-score rms of 4ADH is only ∼8% higher than that of 8ADH (1.22 *versus* 1.13; see Table 5), but we find that the replacement structure contains 30% less atoms with absolute volume *Z*-scores >3. A similar behavior is displayed by the 3CPV/5CPV pair of obsolete and replacement entries for parvalbumin. Here the replacement structure has a better resolution (1.6 Å *versus* 1.85 Å) and lower *R*-factor (0.187 *versus* 0.400). It also has a significantly lower *Z*-score rms than the obsolete entry (1.02 *versus* 1.72), as well as fewer atoms with atomic volume *Z*-scores >3.

In the obsolete structures, backbone atoms (CA, C, N, O and CB) that correspond to bond-length and bond-angle outliers, as defined by PROCHECK, are often associated with volume irregularities. This trend is not evident in the replacement structures.

Discussion

Our study has linked deviations of atomic volumes from their standard values to errors in protein models. To establish this link we defined atom subpopulations and characterized the volume distributions of buried atoms in these subpopulations from a reference set of 64 highly resolved and well-refined protein structures.

In analyzing deviations from standard volumes, it is crucial to distinguish between physically meaningful deviations due to the heterogeneous character of the protein medium, and those caused by model imperfections. This is, however, not an easy task. By revealing correlations between the deviations from the standard atomic volumes and other measures of structure quality, our study suggests that caution should be taken in interpreting trends in atomic volumes in present-day protein structures. This should change in the future, with the rapidly increasing number of protein

crystal structures solved at near atomic resolution (~1 Å), from which it will hopefully be possible to derive more accurate volume distributions.

In addition to more accurate models for the protein moiety, these structures should also yield more reliable positions for a larger number of solvent molecules. This may not only help to characterize the volume of accessible atoms, but should also help improve our description of the volumes of buried atoms.

We found that the mean volumes of buried atoms (defined as those having no surface area accessible to solvent in the water-free atomic coordinates), and as a result, of residues as well, tend to shrink somewhat when water molecules are included in the calculations. Though the reduction in volume is small (1%) when averaging over all atom types, it can be significant for polar atoms where it ranges from 4.7% (OH1; OH groups of Thr, Ser, and Tyr) to 14% (NC3; Lys N^{ζ} and terminal NH_3). These latter large changes may not be significant, however, given the uncertainty associated with the average atomic volumes computed for these atom types, which are so very rarely buried. Nevertheless, these observations call for caution in analyzing trends in volumes of atoms close to the protein surface deduced from analysis of protein crystal structures (Gerstein *et al*., 1995). Atoms close to the protein surface, but defined as buried by the accessible surface area criterion applied to the water-free protein coordinates, may not always be optimally surrounded by other atoms, but become so upon addition of crystallographic water molecules. The effect described above persists when the volume calculations are performed first in the presence of buried water molecules and then in the presence of all crystallographic waters, and is hence not confined to the influence from neglecting buried water molecules in this study.

Lastly, it is important to consider the possible influence of the refinement procedure on the volume deviations. Though volumes are not directly restrained by any of the commonly used refinement protocols, certain protocols, such as XPLOR (Brünger *et al*., 1987), use molecular dynamics procedures. The non-bonded parameters used by these procedures would be expected to influence the volumes and packing of atoms in the crystal structures. To check for such influence, we analyzed the volume deviations in protein models derived by specific refinement procedures, using a much larger protein sample from the most recent release of the PDB. A preliminary analysis of the results indicates that structures refined by XPLOR displayed on the average similar trends in their volume deviations as structures refined by other procedures, with only very few exceptions. For example, the structures refined by TNT (Tronrud *et al*., 1987) appeared to have, on average, smaller volumes than expected. The confirmation of these observations must, however, wait a more detailed analysis, now in progress.

The volume-based assessment of protein structures, described in this study, is implemented in the program PROVE (PROtein Volume Evaluation), which can be accessed as part of the European Biotech structure validation server on the World Wide Web at the following addresses: in Europe: http://biotech.embl-ebi.ac.uk:8400/ and in the US: http://biotech.pdb.bnl.gov:8400/.

Methods

Volume calculations

The calculations of atomic volumes are performed using the classical Voronoi method (Voronoi, 1908) as implemented in the program SurVol (Alard, 1991). In this method, the volume assigned to each atom is defined as the smallest polyhedron created by the set of planes that bisect the vectors connecting an atom's center to those of its neighbors, as illustrated in Figure 9. Atomic volume calculations in proteins are usually performed using modified versions of the original Voronoi method, which take into account the difference in vdw radii between atoms. In some modifications, the position of the dividing planes are determined on the basis of the distance between atoms, the radii of the atoms and whether the atom pair is bonded or not (Richards, 1974, 1985). With this variant, portions of space remain unassigned to any atom, causing what has been referred to as the vertex error, whose magnitude is, however, negligible (Gerstein *et al*., 1995). In other modifications, the dividing planes are positioned only on the basis of the distance between atoms and their radii (Gellatly & Finney, 1982).

One of the advantages of the modified procedures *versus* the classical method is believed to be that they yield a smaller variance in the volumes computed for individual atom types from the protein database. However, a detailed comparison between volumes computed by the classical Voronoi method and by a procedure similar to that of Gellatly & Finney (1982), using a database derived set of atomic radii, did not show this trend consistently (our unpublished results). Furthermore, assigning a consistent set of atomic radii to proteins is a difficult problem, and there is little hope for obtaining an adequate set of atomic radii for all the ligands and co-factors encountered in protein crystal structures. The parameter-free classical Voronoi procedure was therefore considered as best suited for the structure validation approach described here.

When computing the volume of atoms in proteins, problems are encountered in evaluating the volumes of surface atoms. On the protein surface, many neighbors of a protein atom are solvent molecules, of which only a fraction is located by the crystallographer. Though various methods for defining the volumes of surface atoms have been proposed (Finney, 1975; Gellatly & Finney, 1982; Connolly, 1985; Alard, 1991), none has been widely applied, or systematically evaluated. The present analysis therefore considers only buried atoms, defined as atoms whose accessible surface area to solvent (Lee & Richards, 1971) is zero using a probe radius of 1.5 Å. The solvent accessible surface area is computed using an analytical algorithm analogous to that of Connolly (1983), implemented in the program SurVol (Alard, 1991).

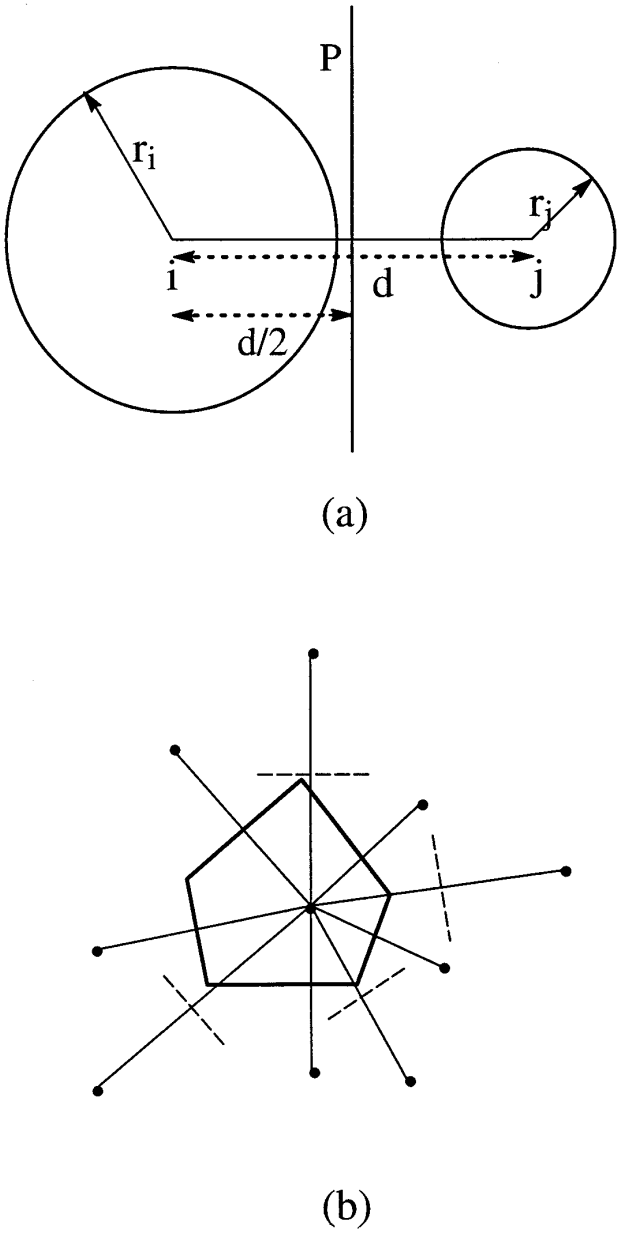


Figure 9. Illustration of Voronoi procedure, for calculating atomic volumes. (a) The classical Voronoi procedure positions the plane P halfway ($d/2$) between the centers of atoms i and j, with radii r_i and r_j, whose center-to-center distance is d. (b) A 2D illustration of the Voronoi polygon (polyhedron in 3D) defining the space (volume) occupied by the central atom when it is surrounded by neighboring atoms in a structure. Vectors are drawn from the central atom to all its neighbors within a given radius, and the planes P perpendicular to those vectors are positioned, as illustrated in (a). The smallest polygon constructed in this way is the Voronoi polygon.

Standard atom and residue volumes

SurVol is used to calculate the Voronoi volumes of atoms in the protein reference set. Water molecules, DNA and hetero group atoms were excluded from volume and surface calculations, as were hydrogen atoms. Protein atoms lining cavities within the structure, such as those created by excluding these groups, or cavities that are empty even when these groups are included, are identified by SurVol and treated as surface atoms and their volumes are not computed. Mean volumes of buried atoms and standard deviations are calculated for atoms grouped according to their atom type, defined by their residue, IUPAC code (IUPAC-IUB, 1969 (1970)) and chemical type. Standard volume ranges are computed only for completely buried atoms, as stated above.

Residue volumes are calculated by summing the volumes of their component atoms.

Measures of volume irregularity

Volume irregularities are measured only for buried atoms, defined as stated above. In addition, buried atoms corresponding to atom types for which there are less than 20 observations in our protein reference set (nearly exclusively buried polar atoms), are not analyzed. For each of the analyzed atoms, a Z-score is calculated, representing the number of standard deviations away its volume is from the mean volume of the atoms having the same atom type:

$$Z\ score_i = \frac{[V_i^k - \overline{V^k}]}{\sigma^k}$$

where V_i^k designates the atomic volume of atom i, having atom type k, calculated using SurVol. $\overline{V^k}$ denotes the mean volume of buried atoms with the same atom type k, and σ^k denotes its associated standard deviation. A negative Z-score means that the atom has a smaller than average volume, whereas a positive score indicates that an atom has a larger than average volume. The expected average Z-score is zero. The Z-score rms deviation from ideality is used as a global measure of departure from the expected behavior in a given set of N atoms, which can be all the atoms of a given protein structure, or atoms with specific attributes, such as the same B-factor range:

$$Z\ score\ rms = \sqrt{\frac{\sum_{i=1}^{N} [Z\ score_i]^2}{N}}$$

Protein structure data sets

The atomic coordinates of 64 high resolution protein structures (Table 4), obtained from the Brookhaven Protein Data Bank (Bernstein *et al*., 1977), are used as the reference ensemble from which the normal volume ranges are derived. These structures were chosen because they had been refined at a resolution of 2 Å or better and to an R-factor of $\leqslant 0.20$. Furthermore, they include representatives of different fold families as described by Orengo (1993). All these structures also contain the atomic coordinates of water molecules whose position was determined by the crystallographers. Only atoms with 100% occupancy were used for calculation of standard volume ranges.

The investigation of volume irregularities is performed using the atomic coordinates of 900 protein structures. They represent crystallographic entries in the 1994 release of the PDB, for which the resolution and R-factor were given, and which contained more than 100 buried atoms. In addition to this test set, we also analyzed eight obsolete structures, and their replacements (Table 5). In all analyzed structures, symmetry-related molecules or subunits, and hydrogen atoms, are not considered. When

Table 4. The 64 protein structures used as references set

PDB code	Protein	Resolution	*R*-factor
1bbp	Bilin binding protein	2.00	0.200
1cob	Superoxide dismutase	2.00	0.176
1csc	Citrate synthase	1.70	0.188
1cse	Serine proteinase-inhibitor	1.20	0.178
1ctf	Ribosomal protein L7/L12	1.70	0.174
1fkf	FK506 Binding protein	1.70	0.170
1fxd	Redoxin II	1.70	0.157
1gd1	Glyceraldehyde dehydrogenase	1.80	0.177
1gp1	Glutathione peroxidase	2.00	0.171
1hoe	Alpha-amylase inhibitor	2.00	0.199
1ifc	Lipid-binding protein	1.19	0.169
1mba	Myoglobin	1.60	0.193
1mbc	Myoglobin (carbonmonoxy)	1.50	0.171
1paz	Cuproprotein-pseudoazurin	1.55	0.180
1pii	Isomerase and synthase	2.00	0.173
1r69	434 Repressor (N-terminal)	2.00	0.193
1rbp	Retinol binding protein	2.00	0.181
1rnh	Ribonuclease H	2.00	0.198
1rop	ROP:COL* E1 Repressor of primer	1.70	0.182
1snc	Staphylococcal nuclease	1.65	0.161
1tgs	Trypsinogen (complex)	1.80	0.186
1thb	Hemoglobin (T state)	1.50	0.196
1trb	Oxidoreductase (flavoenzyme)	2.00	0.177
1ubq	Chromosomal protein: ubiquitin	1.80	0.176
2alp	Alpha-lytic protease	1.70	0.131
2aza	Electron transport protein: azurin	1.80	0.157
2ca2	Carbonic anhydrase	1.90	0.176
2cdv	Heme protein: cytochrome *c*3	1.80	0.176
2ci2	Chymotrypsin inhibitor 2	2.00	0.198
2er7	Hydrolase: endothiapepsin	1.60	0.142
2fb4	Immunoglobulin FAB	1.90	0.189
2fcr	Flavodoxin	1.80	0.188
2fx2	Flavodoxin	1.90	0.170
2gbp	Glucose binding protein	1.90	0.146
2ovo	Ovomucoid (pheasant)	1.50	0.199
2rhe	Immunoglobulin	1.60	0.149
2rsp	Hydrolase: virus protease	2.00	0.144
2sar	Ribonuclease SA	1.80	0.175
2scp	Sarcoplasmic CA binding protein	2.00	0.180
2sga	Proteinase A	1.50	0.126
2sic	Subtilisin	1.80	0.177
2trx	Thioredoxin	1.68	0.165
2tsc	Thymidylate synthase	1.97	0.180
2wrp	DNA binding regulatory protein	1.65	0.180
3blm	Beta-lactamase	2.00	0.163
3chy	Signal transduction protein: CHE*Y	1.66	0.151
3ebx	Erabutoxin B	1.40	0.176
3grs	Glutathione reductase	1.54	0.186
31zm	Lysozyme	1.70	0.157
4bp2	Prophospholipase A2	1.60	0.190
4cla	Acetyltransferase	2.00	0.157
4enl	Lyase: enolase	1.90	0.149
4icb	Calcium-binding protein	1.60	0.188
4ptp	Beta trypsin (dip inhibited)	1.34	0.171
5p21	Oncogene protein C-H-RAS	1.35	0.196
5rub	Lyase: rubisco	1.70	0.180
6tmn	Thermolysin (complex)	1.60	0.171
6xia	D-Xylose isomerase	1.65	0.141
7aat	Aminotransferase	1.90	0.166
8abp	L-Arabinose-binding protein	1.49	0.175
8acn	Lyase-aconitase	2.00	0.161
8dfr	Dihydrofolate reductase	1.70	0.188
9pap	Papain	1.65	0.161
9rnt	Ribonuclease T1 complex	1.50	0.143
9wga	Agglutinin (isolectin2)	1.80	0.175

Column 1 gives the PDB code. A descriptor of the protein as it appears in the PDB header, or compound records, is given in column 2, the resolution at which the structure was determined is in column 3, and its *R*-factor, in column 4.

Table 5. Obsolete and replacement structures

Obsolete				Replacement			
PDB	Resolution	*R*-factor	*Z*-score rms	PDB	Resolution	*R*-factor	*Z*-score rms
3CPV	1.85	0.40	1.72	5CPV	1.6	0.187	1.02
1FBJ	2.6	0.19	1.29	2FBJ	1.95	0.194	1.10
1RN3	1.45	0.26	1.07	3RN3	1.45	0.2233	0.96
1INS	1.5	0.179	1.06	4INS	1.5	0.153	0.95
1MB5	1.8	na	1.26	2MB5	1.8	na	0.93
2MBN	2.0	na	1.18	4MBN	2.0	0.172	1.08
3MBN	2.0	na	1.22	5MBN	2.0	0.179	1.07
4ADH	2.4	0.26	1.22	8ADH	2.4	0.19	1.13

Summary of the data on the eight analyzed obsolete/replacement protein structure pairs. Each row of the Table gives the data on one such pair. Columns 1 and 5 list the PDB code of the obsolete and replacement entry, respectively. The resolution, *R*-factor and *Z*-score rms of the corresponding entries are listed in columns 2 and 6, 3 and 7, and 4 and 8, respectively. The analyzed proteins are: calcium binding parvalbumin (3cpv, 5cpv); immunoglobin Ig* A Fab fragment (1fbj, 2fbj); ribonuclease A (1rn3, 3rn3); insulin (1ins, 4ins); carbon monoxy myoglobin (1mb5, 2mb5); myoglobin(met) (2mbn, 4mbn); myoglobin(deoxy) (3mbn, 5mbn); apo-liver alcohol dehydrogenase (4adh, 8adh).

alternative side-chain conformations were given, the conformation with highest occupancy is used.

Acknowledgements

We thank Alexei Vaguine and other members of the European Consortium on Structure Validation for stimulating discussion. We thank Philippe Alard, the author of the program SurVol, for valuable help with his program. Rob Hooft is thanked for the World Web procedures and we are indebted to Roman Laskowski for his critical reading of the manuscript. The reported work was part of the European BIOTECHNOLOGY project BIO2-CT92-0524 on 3D Macromolecular Structure Validation. We acknowledge support from the Belgian programme of Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture; the Université Libre de Bruxelles fellowship fund, the Fund for Joint Basic Research (Belgium).

References

Alard, P. (1991). PhD thesis dissertation, Calculs de surface et d'énergie dans le domaine des macromolécules. Université Libre de Bruxelles.

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

Brooks, B. R., Bruccoreri, R. E., Olafson, D., States, D., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculation. *J. Comp. Chem.* **4**, 187–217.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). Crystallographic *R*-factor refinement by molecular dynamics. *Science*, **235**, 458–460.

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

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

Connolly, M. L. (1985). Computation of molecular volume. *J. Am. Chem. Soc.* **107**, 1118–1124.

Delhaise, P., Van Belle, D., Bardiaux, M., Alard, P., Hamers, P., Van Cutsem, E. & Wodak, S. (1985). Analysis of data from computer simulations on macromolecules using the CERAM package. *J. Mol. Graph.* **3**, 116–119.

Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structures refinement. *Acta Crystallog. sect. A*, **47**, 392–400.

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

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

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

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**, 611–649.

Hendrickson, W. A. & Konnert, J. H. (1980). In *Computing in Crystallography* (Diamond, R., Ramaseshan S. & Venkatesan K., eds), pp. 1301, Indian Acad. Sci., Bangalore.

Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93–105.

IUPAC-IUB Commission on Biochemical Nomenclature. (1970). Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). *Biochemisty*, **9**, 3471–3479.

Jack, A. & Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and *R*-factor. *Acta Crystallog. sect A*, **34**, 931–935.

Laskowski, R. A., Moss, D. S. & Thornton, J. (1993a). Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231**, 1049–1067.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993b). PROCHECK: a program to

check the stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283–291.
Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
Luthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (1994). Knowledge based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. *Curr Opin. Struct. Biol.* **4**, 731–737.
Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Struct. Funct. Genet.* **12**, 345–364.
Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.
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. (1985). Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol.* **115**, 440–464.
Sippl, M. J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known conformations. *Proteins: Struct. Funct. Genet.* **13**, 258–271.
Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). Occurrence and role of *cis* peptide bonds in protein structures. *J. Mol. Biol.* **214**, 253–260.
Stroud, R. M. & Fauman, E. B. (1995). Significance of structural changes in proteins: expected errors in refined protein structures. *Protein Sci.* **4**, 2392–2404.
Sussman, J. L. (1985). In *Methods in Enzymology* (Wyckoff, H. W. *et al*., eds), vol. 115, pp. 271–303, Academic Press, Florida.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallog. sect. A*, **43**, 489–501.
Voronoi, G. F. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J. Reine Angew. Math.* **134**, 198–287.

Edited by R. Huber

(Received 20 May 1996; received in revised form 19 August 1996; accepted 28 August 1996)