Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The protein content in crystals and packing coefficients in different space groups

Klas M. Andersson[a]* and Sven Hovmöller[b]

[a]Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, and [b]Department of Structural Chemistry, Stockholm University, S-106 91 Stockholm, Sweden

Correspondence e-mail: klas@chem.gla.ac.uk

Received 4 November 1999
Accepted 4 April 2000

A precise way of estimating the packing coefficient, *i.e.* the ratio between the protein and unit-cell volume, or solvent content in protein crystals is given. At present, the solvent content is not given for most proteins in the Protein Data Bank and in many cases where it is given the values are dubious. The mean density of proteins in the crystalline form is around 1.22 g cm^{-3}, not 1.35 g cm^{-3} as usually stated. This is equivalent to 19.5 Å^3 per non-H atom. A statistical investigation of the average protein content and packing coefficient in different space groups is presented. The packing coefficients are generally higher in the most frequently occurring space groups than in the uncommon space groups. There is also a remarkable difference in frequency distribution for enantiomorphous pairs of space groups.

1. Introduction

The packing coefficient of organic molecular crystals is often in the range 65–77% (Kitaigorodskii, 1973). However, in protein crystals the packing coefficient (1 − solvent fraction) is lower. Crick & Kendrew (1968) gave an estimate of 40–60%. This estimate was based on the assumption that proteins have a density of 1.35 g cm^{-3}. Given a revised value for the protein density of 1.22 g cm^{-3} (Andersson & Hovmöller, 1998), the protein volume and hence the packing coefficient will be larger. Most protein-density measurements are carried out in solution, with very low protein concentrations. The extrapolated protein densities in solution are assumed to be equal to the protein densities in the crystalline state, where the protein concentrations are usually very high ($c \simeq 60\%$ by volume). Kim & Kauzmann (1980) found that the protein density in solution increases at very low protein concentrations. This effect does not fully explain the difference between our findings and the solution experiments, but might be a contributing effect. However, in crystallography, protein densities in the solid state are physically more sound than protein densities in solution.

Matthews (1968) stated that ‘the solvent content is most commonly near 43%’. However, there is no information whether this value is a mean value or if there is a high frequency of proteins having this value. Furthermore, he did not discuss the relation between space-group symmetry and packing coefficient. His selection of proteins is rather limited, since there were so few protein structures solved at that time. Only globular proteins were selected and the 116 selected proteins represent only about ten different protein molecules. Thus, we decided to make a more extensive investigation of the packing coefficient in protein crystals.

The five most common of the 65 chiral space groups, $P2_12_12_1$, $P2_1$, $C2$, $P2_12_12$ and $P3_221$, represent nearly 60% of the entries in the Protein Data Bank (PDB) (Teplyakov & Vainstein, 1990) (Table 1). The maximum packing coefficient for perfect spheres is around 74%. Although proteins are often rather globular, the protein content in crystals is nearly always much less than 74%. The purposes of this investigation are to show how the packing coefficient in protein crystals can be estimated and to show how the packing coefficient varies with space-group symmetry. The packing coefficient, together with the volume of the protein, is a very important constraint in, for example, very low resolution phasing of proteins (Andersson & Hovmöller, 1996; Andersson, 1999), solvent flattening, skeletonization and solvent flipping using density modification (Wang, 1985; Zhang & Main, 1990; Cowtan & Main, 1996; Abrahams & Leslie, 1996). The packing coefficient in different space groups is also an important indicator for estimating the number of subunits in the asymmetric unit.

2. Methodology

The proteins for the present study were selected according to the frequency distribution of the space groups in the PDB. For the

Table 1
Frequency distribution of entries in the PDB and packing coefficients of the 12 most common space groups.

7384 entries of proteins with a sequence >15 amino acids (22 March 1999). N is the number of entries used in this study. $\bar{k}$ ± e.s.d. is the average packing coefficient ± the estimated standard deviation of this value.

Space groups	Frequency (%)	N	Range	$\bar{k}$ ± e.s.d.
$P2_12_12_1$	22.87	20	0.453–0.751	0.61 ± 0.02
$P2_1$	13.64	20	0.442–0.803	0.60 ± 0.02
$C2$†	8.94	14	0.432–0.698	0.58 ± 0.02
$P3_221$	6.65	10	0.349–0.696	0.52 ± 0.03
$P2_12_12$	6.15	10	0.429–0.727	0.55 ± 0.03
$P4_32_12$	5.39	10	0.306–0.707	0.51 ± 0.04
$C222_1$	4.39	10	0.392–0.651	0.55 ± 0.03
$P3_121$	3.70	10	0.256–0.741	0.53 ± 0.04
$P1$	2.74	10	0.477–0.783	0.62 ± 0.03
$P4_12_12$	2.73	10	0.302–0.709	0.50 ± 0.03
$I222$	2.45	6	0.365–0.590	0.49 ± 0.03
$P6_122$	1.83	5	0.470–0.574	0.52 ± 0.02
$R3$‡	1.41	5	0.333–0.637	0.50 ± 0.06
Remaining	17.11	40	0.182–0.659	0.46 ± 0.02
k_w§				0.566 ± 0.061

† $A2$, $B2$ and $I2$ included. ‡ $H3$ (hexagonal setting) included. § k_w is the frequency-weighted packing coefficient of all space groups.

two most common space groups, $P2_12_12_1$ and $P2_1$, 20 proteins were selected among the entries by choosing every ith entry, where i is the number of entries divided by 20. For the next most frequent space groups $C2$, $P3_221$, $P2_12_12$, $P4_32_12$, $C222_1$, $P3_121$ and $P1$, ten entries were selected using an analogous selection rule. For $I222$, $P6_122$, $R3$, $I4$ and $P3$, five or six entries were selected. In the remaining 41 very rare space groups all the entries available were selected in order to allow a statistical comparison of the packing coefficients between the different space groups. 181 entries having an amino-acid sequence of >15 residues were selected. Nine very small polypeptides of 15–100 amino acids were included; the other proteins were larger.

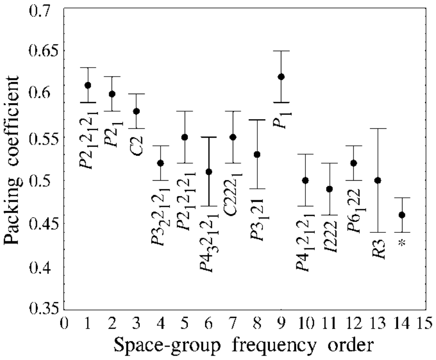

Figure 1
The decline of the packing coefficient *versus* the order of space-group frequency distribution. The star designates the remaining uncommon space groups in this study. The error bars represent the estimated standard deviation (e.s.d.) of the average packing coefficients.

The packing coefficient k was calculated using

$$k = \frac{ZV_{\rm pro}}{V_{\rm cell}} = \frac{19.5ZN^{\rm tot}_{n-H}}{V_{\rm cell}}, \qquad (1)$$

where $V_{\rm pro}$ and $V_{\rm cell}$ are the protein and cell volumes, respectively, Z is the number of protein molecules in the unit cell and $N^{\rm tot}_{n-H}$ is the total number of non-H atoms in the molecule excluding solvent molecules. The protein volume $V_{\rm pro}$ was calculated from the number of non-H atoms in the amino-acid sequence plus the heterogenous atoms and the known mean non-H atomic volume (19.5 Å^3) of proteins (Andersson & Hovmöller, 1998).

3. Results and discussion

In the PDB, solvent content is given for less than half of all entries and for only 71 of the 181 proteins studied here. In most cases, the solvent is overestimated by about 10% (for instance 50% when 45% is more correct, but there are considerable variations). This discrepancy follows from the different density values used for protein crystals *i.e.* 1.35 g cm^{-3} instead of 1.22 g cm^{-3}.

The weighted[1] packing coefficient k_w is 0.566 ± 0.061 for all proteins (Table 1). This value is remarkably similar to the 57% that Matthews (1968) reported, in spite of the revised density value used here.

There is a clear tendency for the most common space groups to have higher packing coefficients. The reason for this might be the high number of rigid-body degrees of freedom (D) in $P2_12_12_1$ and $P2_1$ (Wukovitz & Yeates, 1995). For instance, there are no proteins in space group $P222$ because of the very restricted packing of molecules. A driving force in crystallization is to maximize the number of contacts between molecules and thus also to maximize crystal density and minimize free volume. Void space in crystals is always thermodynamically unfavourable. The most common space groups have the highest packing coefficients and are thus the most thermodynamically favourable. There is a large range of packing coefficients for individual crystals in each space group.

The trend of packing coefficients of different space groups is almost linear with the frequency of entries in the PDB (Fig. 1). However, space group $P1$ is an outlier, having a mean packing coefficient as high as that of $P2_12_12_1$, yet being less common. One reason for this might be that many proteins crystallize in $P1$, but are not structurally determined owing to crystallographic difficulties in this space group. These proteins may have also been crystallized in another space group under different crystallization conditions and thus not reported as $P1$ to *e.g.* the PDB.

Only two water-soluble proteins were found to have packing coefficients of less than 20% (1wip in $P2$ and 1frt in $I2_12_12_1$).

An interesting observation is that the enantiomorphous pairs of space groups differ significantly in frequency distribution. $P3_221$ is about twice as common as its enantiomorphous analogue $P3_121$ (Table 1). This is also the case for the pairs ($P4_32_12$, $P4_12_12$) and ($P6_522$, $P6_122$), where $P4_32_12$ and $P6_122$ are the more common space groups. The reasons for these highly significant differences in frequency between enantiomorphous pairs of space groups are unknown. It remains a very intriguing question.

This study was supported by a grant from the European Commission to the project New Frontiers in Structural Biology.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Andersson, K. M. (1999). *J. Appl. Cryst.* **32**, 530–535.
Andersson, K. M. & Hovmöller, S. (1996). *Acta Cryst.* D**52**, 1174–1180.
Andersson, K. M. & Hovmöller, S. (1998). *Z. Kristallogr.* **213**, 369–373.
Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**52**, 43–48.
Crick, F. H. C. & Kendrew, J. C. (1968). *J. Mol. Biol.* **33**, 491–497.
Kim, K. & Kauzmann, W. (1980). *J. Phys. Chem.* **84**, 163–165.
Kitaigorodskii, A. (1973). *Molecular Crystals and Molecules.* New York: Academic Press.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Teplyakov, A. & Vainstein, B. (1990). *Sov. Phys. Crystallogr.* **35**(3), 414–418.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Wukovitz, S. W. & Yeates, T. O. (1995). *Nature Struct. Biol.* **2**, 1062–1067.
Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst.* A**46**, 377–381.

[1] The mean value is weighted according to the frequency distribution of each space group.