Analysis of solvent content and oligomeric states in protein crystals—does symmetry matter?

MAKSYMILIAN CHRUSZCZ,¹ WOJCIECH POTRZEBOWSKI,^{1,2}
MATTHEW D. ZIMMERMAN,¹ MAREK GRABOWSKI,¹ HEPING ZHENG,¹
PIOTR LASOTA,¹ AND WLADEK MINOR¹

¹Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, Virginia 22908, USA

²Technical University of Lodz, Institute of Physics, 90-924 Lodz, Poland

(RECEIVED November 15, 2007; FINAL REVISION January 24, 2008; ACCEPTED January 29, 2008)

Abstract

A nonredundant set of 9081 protein crystal structures in the Protein Data Bank was used to examine the solvent content, the number of polypeptide chains, and the oligomeric states of proteins in crystals as a function of crystal symmetry (as classified by crystal systems and space groups). It was found that there is a correlation between solvent content and crystal symmetry. Surprisingly, proteins crystallizing in lower symmetry systems have lower solvent content compared to those crystallizing in higher symmetry systems. Nevertheless, there is no universal correlation between solvent content and preferences of macromolecules to crystallize in certain space groups. Crystal symmetry as a function of oligomeric state was examined, where trimers, tetramers, and hexamers were found to prefer to crystallize in systems where the oligomer symmetry could be incorporated in the crystal symmetry. Our analysis also shows that the frequency distribution within the enantiomorphous pairs of space groups does not differ significantly, in contrast to previous reports.

Keywords: solvent content; Matthews coefficient; protein crystals; oligomerization; space group frequency

Supplemental material: see www.proteinscience.org

Water plays an important role in the structure of biomolecules and often influences protein function. Water molecules not only affect protein folding, but also mediate biological processes such as enzymatic reactions and molecular recognition. Information about the fraction of water (solvent) plays a significant role in the X-ray structure determination process. First, knowledge of the solvent content helps to determine the number of molecules in the asymmetric unit (Matthews 1968), which is

crucial in early stages of crystal structure determination. Second, an approximate value of solvent content is needed for significant phase improvement by solvent flattening methods (Wang 1985; Leslie 1987; Abrahams and Leslie 1996), which is necessary to resolve the inherent phase ambiguity in single anomalous diffraction (SAD) experiments. For both SAD and MAD (multi-wavelength anomalous diffraction) (Hendrickson 1991; Hendrickson et al. 1990), phase improvement by solvent flattening is critical for low resolution data (Kirillova et al. 2007), especially when non-crystallographic symmetry cannot be applied.

Matthews (1968) observed that the solvent content in protein crystals ranged from 27% to 65%, with an average of 43%. He also showed that the quantity V_M (the Matthews coefficient, defined as the ratio of the volume of the asymmetric unit to the molecular weight of all

Reprint requests to: Wladek Minor, Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA; e-mail: wladek@iwonka.med.virginia.edu; fax: (434) 982-1616.

Abbreviations: AU, asymmetric unit; V_M , Matthews coefficient.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.073360508>.

protein molecules in the asymmetric unit) is inversely proportional to solvent content. Frequency distributions of V_M for protein crystals in the Protein Data Bank (PDB) (Berman et al. 2002) have been calculated by various studies (Matthews 1968; Andersson and Hovmöller 2000; Kantardjieff and Rupp 2003). While no correlation has been observed between V_M and molecular weight of the protein, a correlation between V_M and resolution of the diffraction data was reported, indicating that protein crystals with less solvent tend to diffract better (Kantardjieff and Rupp 2003).

Andersson and Hovmöller (2000) analyzed solvent content in the 12 most frequent space groups in the PDB and observed that the more frequent the space group, the smaller the average solvent content. Therefore it appears that solvent content variation might help answer the long-standing question of why proteins and other chemical compounds prefer to crystallize in a select few space groups.

Kitajgorodskij (1961), who analyzed crystals of organic compounds, pointed out that some space groups are preferred because they allow a greater likelihood of forming closely packed structures. In most cases, the categorization proposed by Kitajgorodskij (1961) successfully explains the relative frequencies of space groups of organic crystals. Some anomalies in this theory of close packing were discussed by Wilson (1993). The distribution of space group frequencies for crystals of organic molecules was also analyzed in detail by Filippini and Gavezzotti (1992), who formulated a theory of “good” and “bad” symmetry operators. Their method estimates the relative importance of symmetry operators by measuring the contribution of each operator to the total packing energy.

To explain the space group distribution of monomeric protein crystals, an entropic model was presented by Wukovitz and Yeates (1995). This model was used for simulation of protein crystal nucleation, and the distribution of simulated nuclei was nearly the same as the distribution of macroscopic crystals (Pellegrini et al. 1997) observed in monomeric proteins. Currently the entropic model and mathematic dimensionality introduced by Wukovitz and Yeates (1995) provide the best explanation for nonuniformity of space group symmetries in protein crystallography, despite the fact that it does not assume anything about the properties of the molecules. The analysis is more difficult when oligomeric structures are included. In many cases internal symmetry of an oligomer is utilized as part of the crystal symmetry, and as reported recently (Banatao et al. 2006), inducing oligomerization can facilitate crystallization. Regrettably, none of the existing theories attempting to explain the frequency distribution of space groups can describe all of the data. In particular, there is no common theory equally

applicable to inorganic, organic, and macromolecular crystals.

Some of these theories have been evaluated by statistical analysis of structural databases. Padmaja et al. (1990) calculated the distribution of the compounds listed in the Cambridge Structural Database (Allen 2002) that crystallized in the 65 chiral space groups (space groups without operators of the second kind). Calculations narrowed to nucleosides, nucleotides, amino acids, and peptides showed that almost 80% of these compounds crystallized and determined in $P2_12_12_1$ (50.3%) and $P2_1$ (29.0%) space groups. The next most popular space groups were $P1$, $C2$, $P2_12_12$, $P4_12_12$, $C222_1$, $P4_1$, $R3$, $P4_32_12$, $P3_12_1$, $P3_1$, $P6_1$, $P6_3$, $P2_13$, $P6_5$, and $P4_3$ (in order of decreasing frequency). Padmaja et al. (1990) also analyzed deposits in the PDB that contained at that time 208 nonredundant protein structures (synthetic oligopeptides and viral structures were excluded; only one entry from sets of isomorphous structures was considered) and showed that about half of the deposits belonged to three space groups: $P2_12_12_1$ (26.9%), $C2$ (12.9%), and $P2_1$ (10.6%), followed by (in order of decreasing frequency) $P3_12_1$, $C222_1$, $P3_22_1$, $P4_12_12$, $P4_32_12$, $P2_12_12$, and $P1$ (Padmaja et al. 1990). Remarkably, a decade and 7000 structures later, the list of most frequent space groups in the PDB had not changed significantly. Andersson and Hovmöller (2000) examined 7384 crystal protein entries in the PDB as of March 1999 (no provisions were made for dealing with redundant structures) and found the three most frequent groups were still $P2_12_12_1$ (22.9%), $P2_1$ (13.6%), and $C2$ (8.9%). This study also reported that the frequency distribution of enantiomorphous pairs of space groups differed significantly.

In this work, we reexamine the solvent content distribution of protein crystals, based on a nonredundant subset of the PDB as of February 2007. In particular, we analyze correlations between solvent content and crystal symmetry (taking into account crystal system, cell type, and space group). We also examine whether the solvent content of a protein crystal depends on its oligomeric state or on the number of chains in the asymmetric unit. In addition, we investigate whether the solvent content is correlated with the frequency distribution of specific space groups and Bravais lattices among protein crystal structures in the PDB.

Results

Matthews coefficient

For the nonredundant PDB data set containing 9081 proteins considered in this article, the mean Matthews coefficient was 2.68 (0.78) Å³/Da and the median was 2.48 Å³/Da. This is in agreement with the results of

Kantardjieff and Rupp (2003), who reported a mean V_M of $2.69 \text{ \AA}^3/\text{Da}$ and a median of $2.52 \text{ \AA}^3/\text{Da}$ for their set of 10,471 protein crystal structures.

We analyzed correlations between the Matthews coefficient (solvent content) distribution and the crystal system, cell type, and space group. The mean Matthews coefficient varies as a function of crystal system (Table 1; Fig. 1). By treating the seven crystal systems as a set of equally spaced categories, ordered by increasing degree of symmetry, a linear regression of mean V_M weighted by standard deviation is statistically significant (Fig. 2). Thus there is a correlation between the mean V_M and symmetry, where higher symmetry correlates to higher mean V_M values. By degree of symmetry we use the L -value described by Wukovitz and Yeates (1995), where L is a number of independent parameters for describing the unit cell (for example, $L = 1$ for the cubic system and $L = 6$ for the triclinic system). As the L value decreases, the "degree of symmetry" increases. In the case of trigonal, tetragonal, and hexagonal systems, which have the same value of $L(=2)$, the degree of fold symmetry about the main rotational axis for each system was used as an additional criterion for arranging the systems. For example, the trigonal system is characterized by a threefold axis and the tetragonal system by a fourfold axis, so the tetragonal system was considered to have a higher degree of symmetry. The mean V_M also varies by cell type (Table 1). Primitive and C -centered unit cells have a similar mean V_M , while $R(H)$, I - and F -centered cells have higher mean values of V_M with the maximum value for F -centered cells.

For Bravais lattices, the mean V_M in the orthorhombic and cubic systems depends on the lattice type (Table 2).

In orthorhombic crystals the mean V_M is about 10% higher in F than in P lattices. Similar differences are observed between the primitive cubic (cP) and face-centered cubic (cF) lattices. Table 3 lists the space groups with more than 40 representatives in the test set studied. Interestingly, the difference in mean V_M between most of those 28 space groups is not larger than 10%, in most cases within a standard deviation. Only space groups $I4$, $I4_{122}$, $I4_122$, and $P6_322$ have a significantly higher mean V_M .

We analyzed the mean values of V_M as a function of protein chains in the asymmetric unit (AU), as shown in Table 4. There is no significant correlation between number of protein chains in the AU and the mean V_M . However, structures with one, two, or four molecules in the AU have a slightly lower mean solvent content than others. There is also no observed relationship between mean V_M and protein oligomeric state (Table 5).

Symmetry of protein crystals

Distributions of space groups in the nonredundant set of crystal structures are shown in Table 6 and in supplemental Table S1. The most frequent space groups are $P2_12_12_1$, assigned to about 21% of the structures, followed by $P2_1$ and $C2$, observed in 15.6% and 11.0% of the structures, respectively. The 28 space groups listed in Table 3 represent 93% of all the crystal structures in the test set.

Results obtained from our nonredundant set of proteins show only very small differences in the distribution of protein crystal structures between enantiomorphous pairs

Table 1. Mean, median, and mode Matthews coefficients and solvent contents of protein structures in the test set as a function of crystal system or cell centering type

Sample	Matthews coefficient [$\text{\AA}^3/\text{Da}$]			Solvent content	
	Mean	Median	Mode	Mean	Sample size
All	2.68 (0.78)	2.48	2.31	0.51 (0.10)	9081
Crystal system					
Triclinic	2.40 (0.52)	2.31	2.26	0.48 (0.08)	346
Monoclinic	2.43 (0.48)	2.35	2.21	0.48 (0.09)	2295
Orthorhombic	2.51 (0.57)	2.39	2.34	0.49 (0.09)	3213
Tetragonal	3.04 (1.01)	2.75	2.58	0.56 (0.11)	1210
Trigonal	2.94 (0.93)	2.70	2.56	0.55 (0.11)	1046
Hexagonal	3.08 (0.94)	2.88	2.45	0.57 (0.11)	765
Cubic	3.44 (1.17)	3.07	2.95	0.61 (0.10)	206
Cell type					
P	2.62 (0.70)	2.45	2.26	0.51 (0.10)	6851
I	3.33 (1.22)	2.95	2.71	0.59 (0.11)	559
C, A, B	2.59 (0.60)	2.45	2.30	0.51 (0.09)	1365
F	3.32 (0.95)	3.08	2.95	0.60 (0.10)	50
R, H	3.03 (1.24)	2.68	2.44	0.55 (0.11)	256

Standard deviations for the means are given in parentheses.

Figure 1. Frequency distribution for Matthews coefficient in (A) triclinic, (B) tetragonal, and (C) cubic systems and (D) the whole nonredundant set. Arrows indicate the median values of the Matthews coefficient for the selected crystal systems or for the whole nonredundant set. The histograms show data with V_M in the 1.5–5.0 $\text{\AA}^3/\text{Da}$ range. The last bar on each histogram includes also all structures with V_M values above 5.0 $\text{\AA}^3/\text{Da}$. The results reported in the tables are calculated using all data.

of space groups, contrary to previous reports (Andersson and Hovmöller 2000). Only for the pair of $P6_122/P6_522$ is there a significant difference in frequency seen, as $P6_522$ is about 15% more common (Table 6; supplemental Table S1).

Most proteins from the test set crystallized in the orthorhombic and the monoclinic systems (Table 1). More crystals belong to the tetragonal system than the trigonal system, but if we consider crystal families, the hexagonal family (comprising the hexagonal and trigonal systems) is more “popular” than tetragonal, representing almost 20% of all of the crystals in the set. The cubic system is represented by the smallest number of structures, and the triclinic system also has relatively few representatives.

More than 75% of protein structures from the test set crystallized in the primitive lattice. The *C*-face-centered lattice is the second most common. *I*-, *F*-centered, and *R*(*H*) lattices are observed in less than 10% of the structures of the test set. The *F*-centered lattice is very unusual—it is represented by only 50 structures in the test set. The primitive orthorhombic lattice is the most common Bravais lattice type in the test set and is observed in 27% of the protein structures. Primitive monoclinic cells are also very popular. Over 43% of all structures analyzed in this paper have symmetry described by the primitive orthorhombic (*oP*) and primitive monoclinic (*mP*) lattices. The primitive tetragonal (*tP*), monoclinic base-centered (*mC*), primitive trigonal, and primitive hexagonal (*hP*) lattices are assigned to 10.5%, 9.4%, 8.7%, and 8.4% of the crystals, respectively.

The relative frequency distribution by crystal system was considered as a function of the oligomeric state of the AU. For the sets of protein structures with monomers or dimers in the AU, the relative frequency distribution by crystal system is similar to that seen for all structures (Fig. 3A). However, for higher-ordered oligomers, distributions differ (Fig. 3B). For trimers and hexamers, the hexagonal and trigonal systems are overrepresented (in other words, seen relatively more frequently as compared to the monomer distribution). As should be expected, a relatively large percentage of trimers crystallized in the cubic system as well. For tetramers, the tetragonal and orthorhombic systems are overrepresented, and for octamers, the tetragonal crystal system is the most popular (Fig. 3B). Overall, most of the structures (41%) have one polypeptide chain per asymmetric unit, 34% have two chains per AU, and 12% have four chains per AU. Relatively few structures have five polypeptide chains per AU, almost none have seven, and only about 2% of the protein structures contained more than eight chains per AU. Frequency distribution by oligomeric state for the most frequently observed space groups is shown in Table 7.

Figure 2. A categorical linear regression of mean V_M , using the standard deviations of each mean as weights, as a function of crystal system ordered by increasing degree of symmetry ($R^2 = 0.94$, P -value = 0.0002). The error bars represent the standard deviation for the mean of each category.

Discussion

Removing structures with similar sequences in the PDB to select a nonredundant set of protein structures leads to mean and median values of the Matthews coefficient similar to those obtained by Kantardjieff and Rupp (2003). The similar overall mean and median values of V_M suggest that the possible bias introduced by the clustering algorithm used during the selection process (Li et al. 2001) may be negligible.

Our results for frequencies of most popular space groups in the PDB generally agree with previous studies

(Padmaja et al. 1990; Wukovitz and Yeates 1995; Pellegrini et al. 1997; Andersson and Hovmöller 2000). Although the number of PDB deposits has increased more than 100-fold since the analysis by Padmaja et al. (1990), and relative frequencies of the space groups have changed, we found that the three most popular space groups ($P2_12_12_1$, $C2$, and $P2_1$) still contributed about 50% of our set of nonredundant protein crystal structures.

One of our most interesting observations is the slight but statistically significant overall correlation of mean V_M with the crystal system (Figs. 1, 2). In general, higher symmetry is related to higher solvent content. The mean solvent content in cubic crystals is more than 10% higher than in triclinic, monoclinic, and orthorhombic systems. However, this correlation seems to be unrelated to the frequencies of the crystal systems.

Generally, less frequent space groups in the PDB tend to have relatively higher solvent content. However, it is not a universal correlation: The triclinic, monoclinic, and orthorhombic systems (or just the $P1$, $P2_1$, and $P2_12_12_1$ space groups), have very similar mean Matthews coefficients, but their frequencies in the PDB are very different. Therefore, the solvent content alone cannot be used to explain the frequency distribution of protein crystal structures in the PDB among different crystal systems, cell types, Bravais lattices, or space groups.

The overall frequency distribution of crystal symmetries cannot be explained alone by the fact that crystal forms of higher symmetry are usually simpler to work with in crystallographic data collection and processing (which might explain why triclinic crystals appear to be underrepresented in the PDB). Close packing and entropic

Table 2. Mean, median, and mode Matthews coefficients and solvent contents of protein structures in the test set as a function of Bravais lattice

Bravais lattices	Matthews coefficient [$\text{\AA}^3/\text{Da}$]			Solvent content	
	Mean	Median	Mode	Mean	Sample size
aP	2.40 (0.52)	2.31	2.26	0.48 (0.08)	346
mP	2.38 (0.43)	2.32	2.20	0.47 (0.08)	1439
mC, mA, mB, mI	2.52 (0.54)	2.41	2.36	0.50 (0.09)	856
oP	2.44 (0.52)	2.35	2.24	0.48 (0.09)	2468
oC	2.71 (0.67)	2.54	2.33	0.52 (0.09)	511
oI	2.79 (0.59)	2.67	2.37	0.54 (0.09)	219
oF	2.71 (0.56)	2.57	2.52	0.53 (0.09)	15
tP	2.86 (0.80)	2.67	2.47	0.54 (0.10)	954
tI	3.71 (1.38)	3.24	2.71	0.63 (0.12)	256
hP (trigonal)	2.90 (0.79)	2.71	2.55	0.55 (0.10)	790
hR	3.03 (1.24)	2.68	2.44	0.56 (0.11)	256
hP (hexagonal)	3.08 (0.94)	2.88	2.45	0.57 (0.11)	765
cP	3.25 (0.98)	3.00	2.82	0.59 (0.11)	89
cI	3.60 (1.40)	3.16	2.41	0.62 (0.12)	82
cF	3.58 (0.97)	3.17	3.14	0.63 (0.09)	35

Standard deviations for the means are given in parentheses.

Table 3. Mean, median, and mode Matthews coefficients and solvent contents of the protein structure in the test set as a function of the 28 most frequent space groups

Rank	Space group	Matthews coefficient [$\text{\AA}^3/\text{Da}$]			Solvent content	Sample size
		Mean	Median	Mode	Mean	
1	$P2_12_12_1$	2.43 (0.52)	2.33	2.24	0.48 (0.09)	1915
2	$P2_1$	2.38 (0.42)	2.32	2.10	0.47 (0.08)	1417
3	$C2$	2.52 (0.57)	2.41	2.36	0.50 (0.09)	856
4	$P2_12_12$	2.48 (0.46)	2.40	2.27	0.49 (0.08)	534
5	$C222_1$	2.69 (0.66)	2.53	2.33	0.52 (0.09)	487
6	$P1$	2.40 (0.52)	2.31	2.26	0.48 (0.08)	346
7	$P4_32_12$	2.82 (0.79)	2.60	2.55	0.54 (0.10)	330
8	$P4_12_12$	2.93 (0.80)	2.75	2.47	0.56 (0.10)	324
9	$P3_221$	2.96 (0.84)	2.75	2.55	0.56 (0.10)	311
10	$P3_121$	2.96 (0.80)	2.75	2.53	0.56 (0.10)	298
11	$I222$	2.78 (0.6)	2.66	2.37	0.54 (0.09)	202
12	$P6_122$	3.10 (0.94)	2.84	2.43	0.57 (0.11)	135
13	$R32, H32$	3.16 (1.24)	2.79	2.72	0.57 (0.11)	133
14	$P6_522$	3.07 (0.83)	2.87	2.87	0.57 (0.10)	133
15	$R3, H3$	2.90 (1.23)	2.55	2.46	0.53 (0.11)	123
16	$P6_5$	2.92 (0.65)	2.79	2.65	0.56 (0.10)	116
17	$P6_1$	3.00 (0.82)	2.86	2.63	0.56 (0.10)	95
18	$I4_122$	3.33 (0.89)	3.15	2.70	0.61 (0.09)	79
19	$P6_3$	2.90 (1.13)	2.70	2.48	0.54 (0.11)	78
20	$I422$	2.97 (0.64)	2.9	2.71	0.57 (0.08)	77
21	$P6_322$	3.40 (0.93)	3.15	2.88	0.61 (0.10)	73
22	$P4_22_12$	2.90 (1.02)	2.58	2.58	0.54 (0.10)	68
23	$P4_3$	2.72 (0.70)	2.67	2.80	0.53 (0.09)	64
24	$I4$	5.46 (1.26)	5.31	4.35	0.76 (0.05)	62
25	$P3_1$	2.55 (0.51)	2.45	2.64	0.50 (0.09)	58
26	$P3_2$	2.69 (0.61)	2.63	2.32	0.52 (0.10)	50
27	$P4_1$	2.75 (0.53)	2.61	2.18	0.53 (0.09)	48
28	$P2_13$	2.95 (0.81)	2.82	3.04	0.56 (0.11)	46

Standard deviations for the means are given in parentheses.

theories also cannot fully explain such phenomena—for example, they cannot account for the differences between the $P2_1$ and $P1$ space groups. Although the entropic model still does not explain some details of the frequencies of space group symmetry, it currently provides the best

description of the nonuniformity. It is especially visible upon review of the results presented in Table 3 considering the number of rigid-body degrees of freedom (D) introduced by Wukovitz and Yeates (1995). The space groups for which the D-value are highest ($D = 6$ or $D = 7$) occupy

Table 4. Mean, median, and mode Matthews coefficients and solvent contents in the test set of protein structures, as a function of the number of polypeptide chains per AU

Number of protein chains in AU	Matthews coefficient [$\text{\AA}^3/\text{Da}$]			Solvent content	Sample size
	Mean	Median	Mode	Mean	
1	2.66 (0.82)	2.44	2.31	0.51 (0.11)	3724
2	2.65 (0.70)	2.47	2.34	0.51 (0.10)	3100
3	2.85 (0.93)	2.61	2.65	0.54 (0.11)	457
4	2.62 (0.63)	2.47	2.36	0.51 (0.09)	1068
5	2.81 (0.95)	2.52	2.38	0.52 (0.12)	64
6	2.78 (0.82)	2.52	2.51	0.53 (0.10)	301
7	2.92 (0.74)	2.69	2.29	0.56 (0.10)	20
8	2.73 (0.77)	2.57	2.31	0.53 (0.09)	169
>8	3.02 (1.06)	2.73	2.75	0.56 (0.10)	178

Standard deviations for the means are given in parentheses. The data are not limited to identical sequences, so, for example, a heterodimer present in AU is counted as a structure with two chains.

Table 5. Mean, median, and mode Matthews coefficients and solvent contents in the test set of protein structures as a function of protein oligomeric state

Oligomeric state	Matthews coefficient [$\text{\AA}^3/\text{Da}$]			Solvent content	
	Mean	Median	Mode	Mean	Sample size
Mono-	2.59 (0.75)	2.40	2.31	0.50 (0.10)	4437
Di-	2.72 (0.73)	2.52	2.44	0.52 (0.10)	3149
Tri-	2.88 (1.05)	2.56	2.36	0.54 (0.12)	412
Tetra-	2.76 (0.77)	2.53	2.30	0.53 (0.10)	688
Penta-	2.68 (0.82)	2.57	2.09	0.51 (0.11)	42
Hexa-	2.79 (0.72)	2.61	2.45	0.54 (0.10)	214
Hepta-	2.88 (0.90)	2.45	2.36	0.54 (0.11)	9
Octa-	3.43 (1.25)	2.91	2.71	0.60 (0.11)	52
> Octa-	3.03 (0.85)	2.77	2.36	0.57 (0.10)	78

Both homo- and hetero-oligomers were considered. Standard deviations for the means are given in parentheses.

the 13 top positions in Table 3, as the frequency of the $P6_522$ ($D = 6$) space group is the same as $R32$ ($D = 5$).

The theory of “good” and “bad” symmetry operators may also be helpful (Filippini and Gavezzotti 1992). If we assume that the hierarchy of symmetry operators established for organic crystals is also valid for protein crystals, the difference in the number of protein crystal structures reported in $P2_1$ and $P1$ may be explained. During nucleation and/or crystal growth, twofold screw axes may contribute a larger part of the packing energy than translation alone, which may be affected by the shape of the molecule and by the internal symmetry of building blocks. As pointed out by Filippini and Gavezzotti (1992),

objects of irregular shape transformed by twofold axes and mirror planes are influenced by bump-to-bump interactions. Inversion centers, screw axes, and glide planes are better for close packing, which are generated by bump-to-hollow interactions.

Translation is also assumed to be a “good” symmetry operator— $P1$ is the sixth most frequent space group (Table 3). Translation is better for formation of close-packed structures than two-, three-, four-, and sixfold axes. This, at first, may not explain the relative frequency of $C2$, which is the third most often observed space group, after $P2_12_12_1$ and $P2_1$. However, the model does apply, because the combination of C -centering and a twofold

Table 6. Space group distribution

Rank	Space group	Frequency	Number of polypeptide chains in AU								
			1	2	3	4	5	6	7	8	>8
1	$P2_12_12_1$	1915	742	702	76	259	4	73	1	33	25
2	$P2_1$	1417	258	598	40	329	9	70	1	48	64
3	$C2$	856	315	277	85	99	14	30	3	11	22
4	$P2_12_12$	534	211	219	23	52	3	12	1	9	4
5	$C222_1$	487	214	157	41	37	7	10	6	6	9
6	$P1$	346	42	126	14	82	4	20	1	31	26
7	$P4_12_12$	324	164	110	17	17	4	8			4
8	$P4_32_12$	330	199	87	18	15		7	1	3	
9	$P3_221$	311	176	91	20	11	2	6	1	3	1
10	$P3_121$	296	181	85	15	8	2	4		1	
56	$P3_212$	9	5	1				2			1
57	$P23$	8	6	1							1
57	$F4_132$	8	7	1							
59	$P422$	5	2	3							
59	$P432$	5	3	1							1
61	$P222$	2	1		1						
61	$P4_2$	2	1				1				
61	$P4_222$	2	1	1							
64	$P312$	1	1								
65	$P4_232$	0									

The 10 most and 10 least frequent space groups in the set of protein crystal structures analyzed, as a function of the number of polypeptide chains per AU. Statistics for all space groups are shown in the supplemental material (Table S1).

Figure 3. Frequency distribution for different homo-oligomeric assemblies is the crystal system. (A) Monoclinic and orthorhombic systems. (B) Triclinic, tetragonal, trigonal, hexagonal, and cubic systems.

axis yields a twofold screw axis. The results summarized in Table 6 indicate that space groups containing only ordinary crystallographic axes (two-, three-, four-, and sixfold) without screw axes are seldom observed. A

similar observation was made by Wilson (1988) for small molecular crystals. Brock (1996) and Pidcock et al. (2003) also noticed that groups with three-, four-, and sixfold rotation axes occur mainly when molecules or ions are located in special positions on the axes. Based on the space group frequencies observed here, it is possible to classify four- and sixfold screw axes according to their level of frequency, yielding the ordering $4_1 = 4_3 > 4_2$ and $6_1 = 6_5 > 6_3 > 6_2 = 6_4$, respectively. It is interesting that space groups with a symmetry composed of a combination of a three-, four-, or sixfold screw axes with twofold axes appear much more frequently than equivalent space groups with screw axes alone (though this trend is not seen in the tetragonal crystal system). For example, $P6_122$ is found in more structures than $P6_1$. It could be explained assuming that crystallographic twofold axes are coincident with the twofold axes that form protein oligomers. It was recently pointed out that proteins forming oligomers tend to crystallize better than monomeric proteins (Banatao et al. 2006).

High solvent contents are characteristic for protein crystals. Moreover, protein crystals also tend to have more molecules in the AU than do small molecule crystals. A single polypeptide chain in the AU (Table 4) is observed in 41% of the protein structures, while 92% of small molecule crystals have one (or fewer) formula units per AU (Padmaja et al. 1990). Newer reports (Steiner 2000) show that the relative frequencies of the number of molecules in the AU are strongly dependent on crystal system and space group. Steiner reports that organic crystal structures with more than one molecule in the asymmetric unit make up 10% of all organic crystal structures. For steroids, nucleosides, and nucleotides this value reaches 20%. The last column in Table 4 shows that even numbers of polypeptide chains in the AU are preferred (except for single-chain structures). The analysis also shows that if only structures with two or more

Table 7. Frequency distribution by oligomeric state

Rank	Space group	Oligomeric state							
		Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-	Septa-	Octa-
1	$P2_12_12_1$	1171	558 (480)	51 (44)	95 (83)	6 (6)	30 (23)		1 (1)
2	$P2_1$	752	494 (447)	30 (23)	86 (71)	11 (10)	23 (19)		6 (4)
3	$C2$	404	326 (290)	23 (19)	64 (59)	5 (5)	18 (16)	3 (3)	4 (3)
4	$P2_12_12$	238	213 (193)	6 (5)	60 (58)	3 (3)	7 (7)	1 (1)	3 (3)
5	$C222_1$	183	214 (188)	11 (11)	53 (49)	4 (4)	13 (13)	1 (1)	1 (1)
6	$P1$	183	108 (98)	13 (12)	27 (12)	4 (4)	4 (4)	2 (2)	1 (1)
7	$P4_12_12$	156	138 (123)	7 (6)	14 (14)	2 (2)	5 (3)		1 (1)
8	$P4_32_12$	170	124 (109)	7 (6)	19 (16)		8 (7)	1 (1)	
9	$P3_221$	168	108 (95)	4 (4)	20 (16)	1 (1)	8 (8)		1 (1)
10	$P3_121$	137	126 (108)	4 (4)	24 (23)		5 (5)		

The 10 most frequently observed space groups in the analyzed set of protein crystal structures. The frequencies reported are for both homo- and hetero-oligomeric assemblies; frequencies for only homo-oligomeric assemblies are shown in parentheses.

chains in the AU are considered, the frequency of structures with an even number of polypeptide chains per AU (two, four, or six) is higher than the frequency of structures with an odd number of polypeptides per AU (three, five, or seven). Structures with five or seven polypeptide chains per AU are especially rare. Surprisingly, more detailed study reveals that for crystal structures described in certain space groups, especially $P1$, $P2_1$, $P2_12_12$, $P4_3$, $I4$, $P3$, $P3_1$, $P3_2$, and $R3$, two molecules in AU are more often observed than one.

Recently, an interesting approach to protein crystallization by synthetic symmetrization was proposed (Banatao et al. 2006). The distribution of different oligomeric forms in particular space groups (Table 7) and crystal systems (Fig. 3) confirms previous observations showing that internal symmetry of the oligomer could be an advantage during crystallization and such symmetry could be utilized as part of the crystal symmetry. This is observed for the $C222$, $C222_1$, $I222$, $F222$, $I4_122$, and $P6_322$ space groups, where more crystal structures are composed of dimers rather than monomers. Similarly, for trimeric proteins in space groups $P6_3$ and $P2_13$ trimers are observed more often than other oligomeric states. The relative distributions of monomers and dimers in the crystal systems (Fig. 3A) are very similar, but in the case of higher oligomeric assemblies (Fig. 3B) the distribution is strongly influenced by the fact that the internal symmetry of the oligomer may be adopted by crystal symmetry. The most striking examples of such behavior are trimers in trigonal system and tetramers in orthorhombic systems. Another example is the relative absence of pentamers and heptamers, which could be attributed to the inability of these assemblies to be incorporated into crystal symmetry.

Andersson and Hovmöller (2000) reported that frequency distributions of enantiomorphous pairs of space groups differ significantly. They reported that $P3_121$ was about twice as common as $P3_221$ and that this was also the case for the pairs of $P4_12_12/P4_32_12$ and $P6_122/P6_522$. However, such a conclusion appears to be an artifact of not using a nonredundant data set. For example, in the PDB as of September 2007, 1602 structures were solved in $P4_32_12$, versus 1365 structures solved in $P4_12_12$. However, 143 of the $P4_32_12$ structures were all of tetragonal lysozyme. In contrast, our analysis, which took into account only crystals of nonredundant proteins in the PDB, showed only minor differences in frequencies between enantiomorphous pairs of space groups.

The distribution of space-group frequencies may be affected by errors in the reported crystal symmetry. As was pointed out for small molecules (Mighell et al. 1983; Baur and Kassner 1992; Herbstein and Marsh 1998; Marsh 2002; Marsh and Spek 2001; Marsh et al. 2002), many structures are reported in the wrong symmetry, and

in most cases, this wrong symmetry is unnecessarily low (Baur and Tillmanns 1986). In general, there are two types of mistakes in choosing space groups: incorrect space group and correct system or both incorrect space group and system (Marsh 1995; Marsh and Bernal 1995).

Conclusion

We have shown that proteins crystallizing in lower symmetry tend to have a lower solvent content compared to those crystallizing in higher symmetry systems, but the solvent content is not correlated with the distribution of the frequency of crystal structures among different Bravais lattices or specific space groups. We also found that some oligomeric assemblies, like trimers, tetramers, and hexamers, prefer to crystallize in systems where the oligomer symmetry can be incorporated in the crystal symmetry. Moreover, the frequency distributions of enantiomorphous pairs of space groups do not differ significantly. The space group frequencies obtained after analysis of a large nonredundant set of proteins agrees surprisingly well with results obtained by Wukovitz and Yeates (1995) and may be treated as an additional validation of their model.

Materials and Methods

A nonredundant set of protein structures determined by X-ray crystallography was obtained based on the PDB as of February 2007, by clustering proteins at 90% sequence identity (Holm and Sander 1998; Li et al. 2001). Viruses and protein–DNA or protein–RNA complexes were excluded, but protein–protein complexes were included. Our final test set contained 9081 structures, where no structure had greater than 90% sequence identity to any other structure in the set. Use of a clustering algorithm (Li et al. 2001) eliminated bias related to redundant structures in the PDB; however, it might have introduced a slight bias of its own. First, chains of 20 amino acids or less were discarded. Second, the algorithm used resolution and R-value as indicators of the quality of structures to select cluster representatives. Therefore, the test group contained the structures of the highest-resolution crystals in each cluster, and resolution has been observed to be a significant discriminator of V_M (Kantardjieff and Rupp 2003).

The volumes of unit cells were calculated using the unit cell parameters in the CRYST1 records from PDB files. Molecular weights of the polypeptide chains in the asymmetric units were calculated using the sequence information in the SEQRES records and the molecular weights of non-polymerized amino acids. This has limited accuracy as the SEQRES records sometimes give longer sequences than the real sequence of the chains crystallized. Only standard residues were considered: Selenomethionines were treated as methionines. The Matthews coefficient and solvent content were calculated using Matthews' original assumptions (Matthews 1968). Only standard space groups were considered, e.g., structures described in $A2$, $B2$, and $I2$ space groups were considered as $C2$.

The oligomeric states of the proteins were assigned using PITA v0.4.3 (Ponstingl et al. 2003). All oligomeric states with a

PITA score higher than 70 were considered to be potentially significant. When there were multiple possible states detected by the PITA program, the oligomeric state with the largest number of oligomers was used if its score was 80% or greater of the highest returned PITA score.

Acknowledgments

The authors thank Anna Gawlicka-Chruszcz and Zbigniew Dauter for helpful discussions. The work was supported by NIH grants GM74942 and GM53163.

References

- Abrahams, J.P. and Leslie, A.G. 1996. Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr. D Biol Crystallogr.* **52**: 30–42.
- Allen, F.H. 2002. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr. B* **58**: 380–388.
- Andersson, K.M. and Hovmöller, S. 2000. The protein content in crystals and packing coefficients in different space groups. *Acta Crystallogr. D Biol Crystallogr.* **56**: 789–790.
- Banatao, D.R., Cascio, D., Crowley, C.S., Fleissner, M.R., Tienson, H.L., and Yeates, T.O. 2006. An approach to crystallizing proteins by synthetic symmetrization. *Proc. Natl. Acad. Sci.* **103**: 16230–16235.
- Baur, W.H. and Kassner, D. 1992. The perils of Cc—comparing the frequencies of falsely assigned space groups with their general population. *Acta Crystallogr. B* **48**: 356–369.
- Baur, W.H. and Tillmanns, E. 1986. How to avoid unnecessarily low symmetry in crystal-structure determinations. *Acta Crystallogr. B* **42**: 95–111.
- Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. 2002. The Protein Data Bank. *Acta Crystallogr. D Biol Crystallogr.* **58**: 899–907.
- Brock, C.P. 1996. Investigations of the systematics of crystal packing using the Cambridge Structural Database. *J. Res. Natl. Inst. Stand. Technol.* **101**: 321–325.
- Filippini, G. and Gavezzotti, A. 1992. A quantitative analysis of the relative importance of symmetry operators in organic molecular crystals. *Acta Crystallogr. B* **48**: 230–234.
- Hendrickson, W.A. 1991. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**: 51–58.
- Hendrickson, W.A., Horton, J.R., and LeMaster, D.M. 1990. Selenomethionyl proteins produced for analysis by multi-wavelength anomalous diffraction (MAD): A vehicle for direct determination of three-dimensional structure. *EMBO J.* **9**: 1665–1672.
- Herbstein, F.H. and Marsh, R.E. 1998. More space-group corrections: From triclinic to centred monoclinic and to rhombohedral; also from *P*1 to *P*-1 and from *C*c to *C*2/*c*. *Acta Crystallogr. B* **54**: 677–686.
- Holm, L. and Sander, C. 1998. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14**: 423–429.
- Kantardjiev, K.A. and Rupp, B. 2003. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. *Protein Sci.* **12**: 1865–1871.
- Kirillova, O., Chruszcz, M., Shumilin, I.A., Skarina, T., Gorodichtchenskaia, E., Cymborowski, M., Savchenko, A., Edwards, A., and Minor, W. 2007. An extremely SAD case: Structure of a putative redox-enzyme maturation protein from *Archaeoglobus fulgidus* at 3.4 Å resolution. *Acta Crystallogr. D Biol Crystallogr.* **63**: 348–354.
- Kitajgorodskij, A.I. 1961. *Organic chemical crystallography*. Consultants Bureau, New York.
- Leslie, A.G.W. 1987. A reciprocal-space method for calculating a molecular envelope using the algorithm of Wang, B.C. *Acta Crystallogr. A* **43**: 134–136.
- Li, W., Jaroszewski, L., and Godzik, A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**: 282–283.
- Marsh, R.E. 1995. Some thoughts on choosing the correct space group. *Acta Crystallogr. B* **51**: 897–907.
- Marsh, R.E. 2002. The space groups of point group *C*3: Some corrections, some comments. *Acta Crystallogr. B* **58**: 893–899.
- Marsh, R.E. and Bernal, I. 1995. More space-group changes. *Acta Crystallogr. B* **51**: 300–307.
- Marsh, R.E. and Spek, A.L. 2001. Use of software to search for higher symmetry: Space group *C*2. *Acta Crystallogr. B* **57**: 800–805.
- Marsh, R.E., Kapon, M., Hu, S.Z., and Herbstein, F.H. 2002. Some 60 new space-group corrections. *Acta Crystallogr. B* **58**: 62–77.
- Matthews, B.W. 1968. Solvent content of protein crystals. *J. Mol. Biol.* **33**: 491–497.
- Mighell, A.D., Himes, V.L., and Rodgers, J.R. 1983. Space-group frequencies for organic compounds. *Acta Crystallogr. A* **39**: 737–740.
- Padmaja, N., Ramakumar, S., and Viswamitra, M.A. 1990. Space-group frequencies of proteins and of organic compounds with more than one formula unit in the asymmetric unit. *Acta Crystallogr. A* **46**: 725–730.
- Pellegrini, M., Wukovitz, S.W., and Yeates, T.O. 1997. Simulation of protein crystal nucleation. *Proteins* **28**: 515–521.
- Pidcock, E., Motherwell, W.D.S., and Cole, J.C. 2003. A database survey of molecular and crystallographic symmetry. *Acta Crystallogr. B* **59**: 634–640.
- Ponstingl, H., Kabir, T., and Thornton, J.M. 2003. Automatic inference of protein quaternary structure from crystals. *J. Appl. Crystallogr.* **36**: 1116–1122.
- Steiner, T. 2000. Frequency of *Z'* values in organic and organometallic crystal structures. *Acta Crystallogr. B* **56**: 673–676.
- Wang, B.C. 1985. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **115**: 90–112.
- Wilson, A.J.C. 1988. Space groups rare for organic structures. 1. Triclinic, monoclinic and orthorhombic crystal classes. *Acta Crystallogr. B* **44**: 715–724.
- Wilson, A.J.C. 1993. Kitajgorodskij categories. *Acta Crystallogr. B* **49**: 210–212.
- Wukovitz, S.W. and Yeates, T.O. 1995. Why protein crystals favour some space groups over others. *Nat. Struct. Biol.* **2**: 1062–1067.