Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Placement of molecules in (not out of) the cell

Zbigniew Dauter

Synchrotron Radiation Research Section,
National Cancer Institute, Argonne National
Laboratory, Argonne, IL 60439, USA

Correspondence e-mail: dauter@anl.gov

Received 27 September 2012
Accepted 29 October 2012

To uniquely describe a crystal structure, it is sufficient to specify the crystal unit cell and symmetry, and describe the unique structural motif which is repeated by the space-group symmetry throughout the whole crystal. It is somewhat arbitrary how such a unique motif can be defined and positioned with respect to the unit-cell origin. As a result of such freedom, some isomorphous structures are presented in the Protein Data Bank in different locations and appear as if they have different atomic coordinates, despite being completely equivalent structurally. This may easily confuse those users of the PDB who are less familiar with crystallographic symmetry transformations. It would therefore be beneficial for the community of PDB users to introduce standard rules for locating crystal structures of macromolecules in the unit cells of various space groups.

Crystals are built from identical unit cells extending in a parallel fashion in three dimensions. Moreover, each unit cell may contain a number of identical structural motifs (*e.g.* individual molecules or their complexes) arranged according to the symmetry of the particular space group. To uniquely specify the crystal structure, it is therefore sufficient to provide the locations of all of the unique atoms within the asymmetric unit of the crystal, *i.e.* the coordinates of all of these atoms with respect to the cell origin. From a purely crystallographic point of view, it does not matter in which asymmetric unit the specified atoms are located, and both constellations presented in Fig. 1 are equally correct.

However, (molecular) crystals contain chemical compounds, and from the point of view of chemistry the situation in Fig. 1(*a*) is dramatically different from that in Fig. 1(*b*). Whereas in the latter the atomic connectivity and architecture of acridine are immediately apparent, the former representation makes little chemical sense. In analogy, if an asymmetric unit contains several molecules forming discrete oligomers, it is more informative to present the individual molecules grouped logically, rather than randomly, as illustrated in Fig. 2, and indeed most illustrations of oligomeric structures are already presented as biologically relevant assemblies on the PDB web pages.

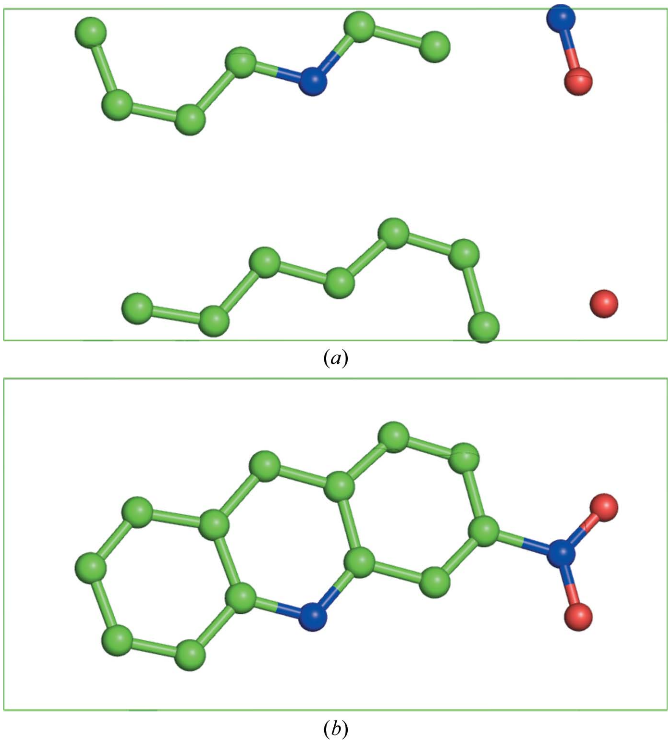
It is worth commenting on the concept of the 'asymmetric unit' (ASU in the following). Intuitively, it is clearly the part of a unit cell which, under the action of all symmetry operations of the space group, reproduces the complete content of the cell and therefore the whole crystal. As long as this requirement is fulfilled, it does not matter what the shape of the ASU is. *International Tables for Crystallography* (2005) contains definitions of the ASU for each space group in the form of a convex parallelepiped (in cubic groups it may be a more complicated polyhedron), but this choice is arbitrary and ASUs may have different shapes. In fact, each molecule or, more strictly, each unique structural motif forms an ASU which may have a quite complicated shape, not necessarily convex. An example of such a construction is the Voronoi (1908) tessellation, which was first applied to protein crystals by Richards (1974).

Apart from crystallographic correctness and chemical sense, the presentation of any macromolecular crystal structure should be logical and as easy to comprehend as possible by other scientists who may be less familiar with the principles of crystallography, *e.g.* biologists interested in the functioning and biochemical properties of a given molecule or complex. In this context, it is meaningful how the structures are presented in the Protein Data Bank (PDB; Berman *et al.*, 2000).

The PDB serves as the repository of macromolecular structures, but it is not responsible for the scientific content of the deposited models. However, it has certain rules concerning the presentation of the atomic models. For example, all solvent water molecules are automatically transformed by symmetry and renamed according to the closest macromolecular chain. The results of this procedure are absolutely equivalent to the original situation and it is meant to make it easier for the results of the structural analysis to be interpreted by people who are less familiar with crystallographic procedures.

The life of noncrystallographers interested in PDB models could be made even easier if some other considerations were also taken into account. Table 1 contains a list of all PDB structures of bovine trypsin complexed with various inhibitors crystallized in the orthorhombic space group $P2_12_12_1$ with similar unit-cell parameters. The list shows the positions of the trypsin molecules in the unit cell of specified dimensions. Any crystallographer would realise that all of these structures are practically isomorphous and therefore the architecture of all of the crystals is equivalent. The unique molecules in different structures are simply transformed by the space-group symmetry or the permissible shift of the cell origin. Non-specialists, however, may be confused and treat these models as structurally different.

In addition to humans, some crystallographic programs are also 'confused' as a result of placing molecules further away from the cell origin. The *CCP*4 program *CONTACT* (Winn *et al.*, 2011) properly identifies six interacting neighboring molecules in most trypsin structures, except for PDB entries 1s0q, 2j9n and 3iti, in which the


(*a*)

(*b*)

Figure 1
Two possible crystallographically equivalent representations of the structure of 3-nitroacridine differing by a half-cell origin shift along the vertical axis and appropriate rearrangement of the atoms. Whereas in (*a*) the molecular structure of this compound is not clear, in (*b*) it is immediately apparent.

Table 1
Average fractional coordinates of all trypsin atoms and unit-cell dimensions (in Å) for each of the nearly isomorphic $P2_12_12_1$ structures of trypsin (in complex with various inhibitors).

PDB code	*x*	*y*	*z*	*a*	*b*	*c*
1auj	0.454	0.372	0.351	55.000	58.400	67.600
1az8	0.453	0.370	0.351	54.800	58.700	67.500
1btx	0.451	0.371	0.348	54.840	58.610	67.470
1bty	0.447	0.373	0.349	54.840	58.610	67.470
1btz	0.452	0.371	0.348	54.840	58.610	67.470
1gi1	0.450	0.375	0.350	54.920	58.830	67.370
1gi2	0.449	0.373	0.351	54.870	58.700	67.330
1hj9	0.457	0.390	0.355	54.055	56.812	66.282
1mtw	0.451	0.362	0.341	54.900	61.000	64.300
1ntp	0.452	0.373	0.351	54.840	58.610	67.470
1ql7	0.450	0.362	0.341	54.060	58.580	63.140
1tio	0.455	0.370	0.351	54.540	58.260	67.410
1tng	0.453	0.371	0.352	54.946	58.424	67.581
1tnh	0.453	0.371	0.351	54.960	58.438	67.562
1tni	0.453	0.371	0.351	54.847	58.550	67.542
1tnj	0.453	0.370	0.352	54.919	58.526	67.535
1tnk	0.453	0.371	0.351	54.948	58.459	67.652
1tnl	0.453	0.371	0.351	54.915	58.501	67.590
1utn	0.455	0.392	0.355	54.150	56.730	66.250
1xuf	0.451	0.370	0.349	54.800	58.700	67.600
1xuk	0.451	0.370	0.348	54.800	58.700	67.600
2by5	0.453	0.374	0.354	54.259	58.337	66.745
2by6	0.450	0.374	0.355	54.262	58.355	66.765
2by7	0.453	0.374	0.354	54.220	58.332	66.751
2by8	0.453	0.374	0.354	54.226	58.347	66.772
2by9	0.453	0.374	0.354	54.245	58.323	66.735
2bya	0.453	0.374	0.354	54.274	58.359	66.782
2ah4	0.455	0.370	0.350	54.446	57.908	66.771
5ptp	0.453	0.372	0.349	54.840	58.610	67.470
1bju	0.045	0.127	0.352	54.840	58.490	67.830
1bjv	0.047	0.127	0.350	54.360	58.200	66.610
1j8a	0.049	0.128	0.350	54.275	58.568	66.141
1k1i	0.049	0.130	0.350	54.904	58.876	67.177
1k1n	0.047	0.130	0.350	54.798	58.273	67.277
1max	0.046	0.127	0.351	54.660	58.480	66.930
1may	0.046	0.128	0.351	54.950	58.550	67.700
1rxp	0.047	0.127	0.351	54.290	58.250	66.680
1tpo	0.046	0.128	0.352	54.890	58.520	67.630
1tpp	0.046	0.129	0.352	54.900	58.500	67.800
1v2o	0.051	0.129	0.353	55.110	58.020	68.520
1v2q	0.049	0.128	0.354	55.050	57.970	68.530
1v2r	0.043	0.116	0.360	55.100	57.780	67.460
1v2t	0.043	0.115	0.359	55.130	58.240	67.810
1v2w	0.048	0.128	0.354	55.010	58.210	68.040
2ptn	0.046	0.129	0.352	54.890	58.520	67.630
3ptb	0.046	0.129	0.351	54.890	58.520	67.630
2blv	0.455	0.124	0.150	54.162	58.253	66.582
2blw	0.455	0.124	0.150	54.162	58.253	66.582
2d8w	0.453	0.126	0.149	54.323	58.197	66.620
1tx7	0.046	0.372	0.649	54.870	58.440	67.520
1tx8	0.046	0.372	0.648	54.795	58.580	67.563
1n6x	0.456	−0.113	0.357	54.209	56.658	66.126
1n6y	0.456	−0.113	0.357	54.241	56.752	66.204
2oxs	0.047	0.128	−0.149	54.756	58.484	67.406
2otv	0.047	−0.372	−0.149	54.831	58.579	67.461
1s0r	0.548	0.628	0.351	54.393	58.424	66.542
1s0q	0.951	0.629	0.151	54.383	58.710	66.427
2j9n	0.450	0.870	0.354	53.924	56.695	66.054
3iti	1.046	0.390	0.140	53.658	56.883	66.808

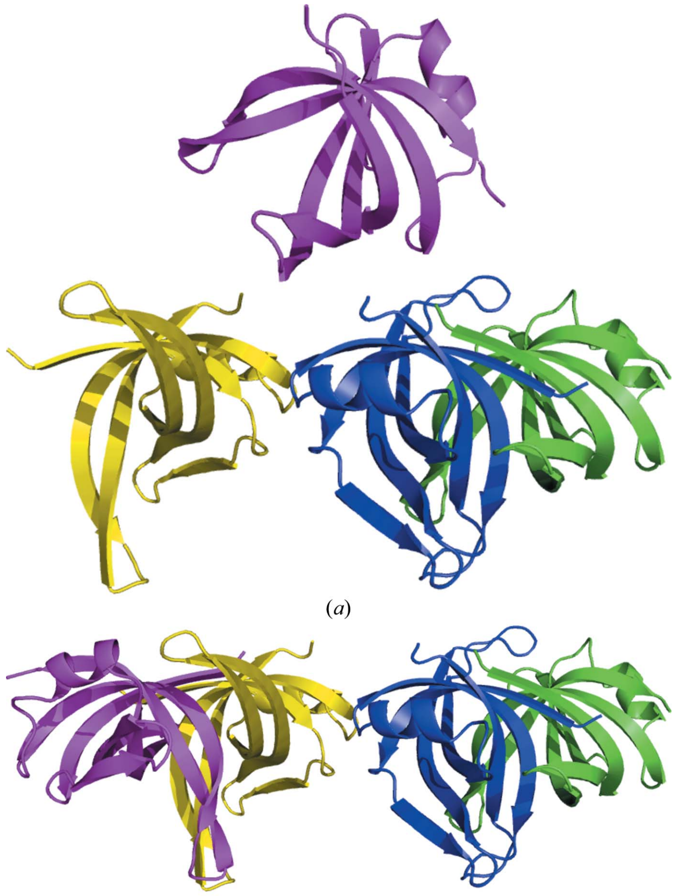
Table 2
Intermolecular contacts identified by the *CCP*4 program *CONTACT* for representative structures of trypsin from Table 1.

PDB code	x	y	z	No. of contacts	Symmetry operations					
5ptp	0.45	0.37	0.35	6	$\frac{1}{2}-x, 1-y, \frac{1}{2}+z$	$\frac{1}{2}-x, 1-y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$-\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$1-x, \frac{1}{2}-y, \frac{1}{2}-z$	$1-x, -\frac{1}{2}-y, \frac{1}{2}-z$
3ptp	0.05	0.13	0.35	6	$\frac{1}{2}-x, -y, \frac{1}{2}+z$	$\frac{1}{2}-x, -y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$-\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$-x, \frac{1}{2}-y, \frac{1}{2}-z$	$-x, -\frac{1}{2}-y, \frac{1}{2}-z$
2d8w	0.45	0.13	0.15	6	$\frac{1}{2}-x, -y, \frac{1}{2}+z$	$\frac{1}{2}-x, -y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{1}{2}-y, -z$	$-\frac{1}{2}+x, \frac{1}{2}-y, -z$	$1-x, \frac{1}{2}-y, \frac{1}{2}-z$	$1-x, -\frac{1}{2}-y, \frac{1}{2}-z$
1tx8	0.05	0.37	0.65	6	$\frac{1}{2}-x, 1-y, \frac{1}{2}+z$	$\frac{1}{2}-x, 1-y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$-\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$1-x, \frac{1}{2}-y, \frac{2}{2}-z$	$1-x, -\frac{1}{2}-y, \frac{3}{2}-z$
1n6y	0.45	−0.13	0.35	6	$\frac{1}{2}-x, -y, \frac{1}{2}+z$	$\frac{1}{2}-x, -y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$-\frac{1}{2}+x, \frac{1}{2}-y, 1-z$	$1-x, \frac{1}{2}-y, \frac{1}{2}-z$	$1-x, -\frac{1}{2}-y, \frac{1}{2}-z$
2oxs	0.05	0.13	−0.15	6	$\frac{1}{2}-x, -y, \frac{1}{2}+z$	$\frac{1}{2}-x, -y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{1}{2}-y, -z$	$-\frac{1}{2}+x, \frac{1}{2}-y, -z$	$-x, \frac{1}{2}-y, -\frac{1}{2}-z$	$-x, -\frac{1}{2}-y, -\frac{1}{2}-z$
2otv	0.05	−0.37	−0.15	6	$\frac{1}{2}-x, -1-y, \frac{1}{2}+z$	$\frac{1}{2}-x, -1-y, -\frac{1}{2}+z$	$\frac{1}{2}+x, -\frac{1}{2}-y, -z$	$-\frac{1}{2}+x, -\frac{1}{2}-y, -z$	$-x, \frac{1}{2}-y, -\frac{1}{2}-z$	$-x, -\frac{1}{2}-y, -\frac{1}{2}-z$
1s0r	0.55	0.63	0.35	6	$\frac{3}{2}-x, 1-y, \frac{1}{2}+z$	$\frac{3}{2}-x, 1-y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{3}{2}-y, 1-z$	$-\frac{1}{2}+x, \frac{3}{2}-y, 1-z$	$1-x, \frac{1}{2}-y, \frac{1}{2}-z$	$1-x, -\frac{1}{2}-y, \frac{1}{2}-z$
1s0q	0.95	0.63	0.15	4	$\frac{3}{2}-x, 1-y, \frac{1}{2}+z$	$\frac{3}{2}-x, 1-y, -\frac{1}{2}+z$	$\frac{1}{2}+x, \frac{3}{2}-y, -z$	$-\frac{1}{2}+x, \frac{3}{2}-y, -z$		
2j9n	0.45	0.87	0.35	4	$\frac{1}{2}+x, \frac{3}{2}-y, 1-z$	$-\frac{1}{2}+x, \frac{3}{2}-y, 1-z$			$1-x, \frac{1}{2}-y, \frac{1}{2}-z$	$1-x, -\frac{1}{2}-y, \frac{1}{2}-z$
3iti	1.05	0.37	0.15	2	$\frac{1}{2}+x, \frac{1}{2}-y, -z$	$-\frac{1}{2}+x, \frac{1}{2}-y, -z$				

protein molecules are located farther from the cell origin (Table 2). As stated in the program documentation,

> the default is to use only single translations (*e.g.* +A, −A, −A+B *etc.*), which works well if the molecule is reasonably positioned within the cell (not outside).

To find the missing neighbors in the 3iti structure, symmetry operations with two-cell translations are required: $2-x, \pm\frac{1}{2}+y, \frac{1}{2}-z$ and $\frac{5}{2}-x, 1-y, \pm\frac{1}{2}+z$. The dimer of the $P2_12_12_1$ structure of polyketide synthase 3hrq is located at fractional coordinates $x = -2.37$, $y = -2.15$, $z = -1.60$, and *CONTACT* is not able to find any intermolecular interactions, whereas in reality there are ten neighboring symmetry-equivalent dimers in this structure.


(*a*)

(*b*)

Figure 2
Four independent protein protomers in the asymmetric unit of the structure 1woc as presented in the PDB (*a*) and after regrouping (*b*), when it becomes apparent that this structure consists of two similar dimers.

Table 3
Placement of molecules in space group $P2_12_12_1$.

All models in this group from the September 2012 content of the PDB are included.

$x, y, z > +3.0$	1
$x, y, z < -2.0$	8
$x, y, z > +2.0$	44
$x, y, z < -1.0$	69
$x, y, z > +1.0$	715
$x, y, z < -0.5$	330
$x, y, z > +0.5$	6189
$x, y, z < 0.0$	6125
$0.0 < x, y, z < +0.5$	5980
All	17188

Taking into account that in space group $P2_12_12_1$ the cell origin can be shifted by half of the unit-cell dimension in any direction, it is always possible to locate the center of the molecule (or, more generally, of the unique structural motif) in the region $-\frac{1}{4} < x, y, z \leq +\frac{1}{4}$. In fact, appropriately selecting one of the four existing orientations of the molecule in the cell, this region can be limited to, for example, $0 \leq x, y < +\frac{1}{4}$, $-\frac{1}{4} < z \leq +\frac{1}{4}$. This is how the structure 2oxs is presented. However, a majority of the $P2_12_12_1$ models in the PDB lie outside this region (Table 3).

Since the atomic coordinates in the PDB are expressed in orthogonal ångström coordinates, their transformation has to proceed through conversion to fractional coordinates. Because of the inevitable limitations of the precision of the stored atomic coordinates and cell dimensions, this procedure will introduce a degree of error that increases with the distance of the model from the cell origin.

In conclusion, it would be beneficial to the wide community of PDB users if all of the structures in this depository could be 'remedied' by shifting their locations to positions as close as possible to the origin of the unit cell.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

International Tables for Crystallography (2005). Vol. *A*, edited by T. Hahn. Heidelberg: Springer.

Richards, F. M. (1974). *J. Mol. Biol.* **82**, 1–14.

Voronoi, G. (1908). *J. Reine Angew. Math.* **134**, 198–287.

Winn, M. D. *et al.*. (2012). *Acta Cryst.* D**67**, 235–242.